# *RELATIONSHIPS BETWEEN PHONETIC PERCEPTUAL AND AUDITORY SPACES FOR FRICATIVES*

**Won CHOO**

Department of Phonetics and Linguistics
University College London

A thesis submitted to
the University of London
for the degree of Doctor of Philosophy

1996

*To my family and Shin*

# *Abstract*

This study investigates the correlations between phonetic, perceptual and auditory spaces of fricatives. These spatial representations are constructed from estimated distances between fricatives in each domain.

The present work is an extension of previous studies with vowels, which showed a close association between the auditory and perceptual spaces. The dimensions obtained were highly related to the formant frequencies of the vowels and their phonetic (articulatory) descriptions. However, these findings could have been anticipated due to inherent similarities in articulatory and acoustic forms of vowels. No corresponding relationships for consonants have yet been established, and it is important to investigate whether such relationships might also hold for consonants, since their acoustic form is not so simply linked to their phonetic (articulatory) form. In this study we relate 'places of articulation' for fricatives with their spectral characteristics and perceptual similarities to investigate any articulatory references to their perception.

A variety of fricative and fricative-like stimuli were examined. Perceptual distances were derived from subjective judgments of the similarities between the fricatives. Auditory distances were obtained from critical bandpass filter banks and distance metrics were applied to model the spectral processing in the auditory periphery. The distances in the perceptual and auditory spaces were analysed using multidimensional scaling in order to test their correlation and how it varied according to the naturalness of the stimulus materials. The relations between the spaces were measured quantitatively by canonical correlation analyses. The acoustic correlates of these spatial dimensions were also identified.

The perceptual dimensions of the most natural fricative stimuli proved to be highly related to both their phonetic (articulatory) and auditory spaces. By demonstrating a link across these domains, the study, therefore, favours perceptual theories of a unified nature, rather than the views based on 'strong articulatory' or 'strong auditory' modes of speech perception.

# *Acknowledgments*

I must express my heartfelt thanks to my supervisor, Mark Huckvale, for introducing me to Multidimensional scaling techniques and to the general ideas contained in the thesis. Without his constant criticism, support, and encouragement, this thesis would not have been possible.

The stimulating atmosphere in Wolfson House provided the opportunity for many spontaneous and lively discussions with members of staff and students; in particular, I thank Richard Baker and Stuart Rosen for helping me through the daunting task of undertaking SAS analysis and interpretation work. Andrew Falkner also helped me with SAS and statistics. In addition, David Cushing was always there to alleviate day-to-day technical problems.

Thanks are also due to other staff members and students of the Department of Phonetics and Linguistics at UCL, for various academic and administrative assistance; in particular, I thank John Wells and Jill House for their help in the earlier stages of my study in the department. Steve Nevard helped with various technical problems in 21 Gordon Square. Also, being together in the Hut, Mercedes Cabrera-Abreu and I shared common difficulties as post-grad students. Phillip Backley proof-read the final version of the thesis.

# *Table of contents*

*Abstract* ..... 3

*Acknowledgments* ..... 4

*Table of contents* ..... 5

*List of figures* ..... 11

*List of tables* ..... 16

*Chapter I. Introduction*

1 Invariance and variability in speech perception ..... 18

2 Methods of investigation ..... 19

3 Study of spatial representations ..... 21

3.0 Introduction ..... 21

3.1 Articulatory/Phonetic spaces ..... 21

3.2 Acoustic/auditory spaces ..... 25

3.3 Perceptual spaces ..... 27

3.4 Spatial representations and modelling of perceptual processes 31

4 Objectives ..... 32

5 Outline of thesis ..... 33

*Chapter II. Previous studies*

1 Introduction ..... 35

2 Spatial relationship ..... 35

2.0 Introduction ..... 35

2.1 Perceptual and physical spaces in vowels ..... 36

2.1.0 Introduction ..... 36

2.1.1 Pols, Kamp & Plomp (1969) ..... 36

2.1.2 Klein, Plomp & Pols (1970) ..... 41

2.2 A study of 'limited' vowel space ..... 43

2.3 Acoustic interpretation of Miller & Nicely's data ..... 48

2.4 Spatial representation of stop consonants ..... 53
3 Modelling of perceptual distance judgments ..... 55
3.1 Distance metrics ..... 55
3.1.0 Introduction ..... 55
3.1.1 Weighted slope metric ..... 55
3.1.2 Weighted negative second differential metric ..... 57
3.2 Some attempts at auditory distance modelling for consonants 59
3.2.0 Introduction ..... 59
3.2.1 Uniform alignment of spectra ..... 59
3.2.2 'Burst length' ..... 61
4 Fricative studies ..... 63
4.0 Introduction ..... 63
4.1 Frication amplitude cue ..... 64
4.2 Spectral formant cue ..... 66
4.3 Vowel quality and formant transition cues ..... 67
4.4 Frication duration cue ..... 68
5 Summary ..... 68

***Chapter III. Preliminary analyses on phonetic, perceptual and auditory spaces of fricative sounds***

1 Introduction and objectives ..... 70
2 Experiment 1: Perceptual analyses of 'natural' fricatives ..... 71
2.0 Introduction ..... 71
2.1 Stimuli ..... 71
2.2 Data collection method ..... 72
2.3 Subjects and procedure ..... 74
2.4 Analyses ..... 76
2.5 Results ..... 78
2.5.1 Metric analyses ..... 78
2.5.2 Nonmetric analyses ..... 82
2.6 Discussion ..... 85

3 Experiment 2: Preliminary analyses on phonetic, perceptual, and auditory spaces of fricative sounds ..... 87
3.0 Introduction ..... 87
3.1 Perceptual analysis ..... 87
3.1.1 stimuli ..... 87
3.1.2 Subjects ..... 89
3.1.3 Perceptual space ..... 89
3.1.3.1 Set 1 ..... 89
3.1.3.2 Set 2 ..... 91
3.2 Auditory analyses ..... 92
3.2.0 Introduction ..... 93
3.2.1 Critical bandpass analyses ..... 93
3.2.2 Non-linear time alignment ..... 96
3.2.3 Distance metrics ..... 96
3.2.4 Auditory space ..... 100
3.3 Relationship between the acoustic and perceptual spaces .. 103
3.4 Discussion ..... 106
4 Experiment 3: Shaped white noises ..... 108
4.0 Introduction ..... 108
4.1 Stimuli design ..... 108
4.2 Subjects ..... 109
4.3 Analyses ..... 109
4.4 Canonical correlations between perceptual and auditory spaces ..... 110
5 Pointers for future experiment design ..... 112

***Chapter IV. Perception tests***

1 Introduction and objectives ..... 114
2 Design of stimuli ..... 115
3 Subjects and procedure ..... 120
4 Results and discusssion ..... 122

4.1 Whole syllable ..... 124
4.1.0 Introduction. ..... 124
4.1.1 Stimulus space for the whole subject group. ..... 124
4.1.2 Subject space. ..... 127
4.1.3 Group A vs. Group B. ..... 128
4.1.4 Summary. ..... 131
4.2 No transition ..... 131
4.2.0 Introduction. ..... 131
4.2.1 Stimulus space for the whole subject group. ..... 132
4.2.2 Subject space. ..... 133
4.2.3 Group A vs. Group B. ..... 134
4.2.4 Summary. ..... 136
4.3 Cut-out ..... 137
4.3.0 Introduction. ..... 137
4.3.1 Group stimulus space. ..... 137
4.3.2 Group A vs. Group B. ..... 138
4.3.3 Summary. ..... 140
4.4 LPC22 ..... 140
4.4.0 Introduction. ..... 140
4.4.1 Group stimulus space. ..... 140
4.4.2 Group A vs. Group B. ..... 141
4.4.3 Summary. ..... 143
4.5 LPC10a ..... 144
4.5.0 Introduction. ..... 144
4.5.1 Group stimulus space. ..... 144
4.5.2 Group A vs. Group B. ..... 145
4.5.3 Summary. ..... 147
4.6 LPC10 ..... 147
4.7 LPC4 ..... 149
4.8 Overview of the results ..... 152

***Chapter V. The relationship between perceptual and auditory spaces***

1 Introduction and objectives ..... 155
2 Euclidean distance metric ..... 156
2.1 Auditory spaces and spectra ..... 156
2.2 Canonical coefficients ..... 163
2.3 Canonical scores ..... 165
3 Slope distance metric ..... 170
3.1 Auditory spaces ..... 170
3.2 Canonical correlations ..... 173
4 Negative second differential (N2D) metric ..... 174
4.1 Auditory spaces ..... 174
4.2 Canonical correlations ..... 177
5 Discussion ..... 178
5.1 Distance metrics ..... 178
5.2 The relationship between perceptual and auditory spaces .. 181

***Chapter VI. Production tests***

1 Introduction and objectives ..... 182
2 Materials: recordings and speakers ..... 182
3 Fricative spectra ..... 182
3.0 Introduction ..... 182
3.1 Initial measurements (duration and loudness) ..... 183
3.2 Auditory spectra ..... 184
3.3 Summary ..... 192
4 Fricative spaces ..... 193
4.0 Introduction ..... 193
4.1 Intra-speaker variations ..... 193
4.2 Inter-speaker variations ..... 198
5 Auditory prototypes and acoustic correlates ..... 199
6 Conclusion ..... 201

***Chapter VII. Conclusion***

1 Introduction ..... 202
2 Main findings ..... 202
2.1 General characteristics of spatial representations ..... 202
2.2 Auditory modelling ..... 203
2.3 Variations in spatial relationship in different stimuli set ... 204
2.4 Summary ..... 205
3 Implications for speech perception theory ..... 206
4 Future developments ..... 207

***References*** ..... 208
***Appendix*** ..... 215
***Addendum*** ..... 222

# *List of figures*

**Figure I-1** The vowel quadrilateral; after Gimson (1989) ..... 22

**Figure I-2** 'Phonetic space' for consonant systems of the world languages .... 24

**Figure I-3** A plot of Mel (F2) against Mel (F1) of cardinal vowels. ..... 26

**Figure I-4** Multidimensional scaling map of the airline distances among ten U.S. cities. ..... 28

**Figure I-5** First two dimensions of MDS solution of the vowels ..... 30

**Figure II-1** Projections of the 11 stimulus points on two perpendicular planes of the three-dimensional physical space. ..... 38

**Figure II-2** Three-dimensional perceptual space of 11 vowel stimuli ..... 39

**Figure II-3** Three most correlating dimensions of perceptual and physical dimensions after rotating to optimal congruence. ..... 42

**Figure II-4** (a) /i-ɪ/ vowels plotted on Bark (F2/F1) plane. (b) Two-dimensional MDS solution of the same vowels. ..... 45

**Figure II-5** (a) The stimulus set /æ-ɑ-ʌ/ plotted on Bark (F2/F1) space (b) The same vowels are plotted in the two-dimensional perceptual space obtained from the MDS analysis. ..... 47

**Figure II-6** The four-dimensional INDSCAL solution of 16 consonants heard in 17 acoustic disturbance conditions.. ..... 49

**Figure II-7** INDCLUS clusters comprised of voiced consonants (I-VI), voiceless consonants (VII-X), and mixed-voiced clusters XI and XII ..... 52

**Figure II-8** A three-dimensional acoustic space of voiced stop consonants /b d g/. Male grand mean (n=10). ..... 53

**Figure II-9** Schematic spectrograms of the symmetrical CVC syllables.. ..... 60

**Figure III-1** An example of stimuli for experiment 1 ..... 72

**Figure III-2** (a) A hypothetical group stimulus space from a 3-way MDS analysis, (b) A corresponding subject space. ..... 77

**Figure III-2** Individual perceptual spaces for (c) Subject 2, (d) Subject 4 ..... 77

**Figure III-3** Fit curve representing the cumulative variance accounted for by metric (INDSCAL) multidimensional scaling analyses ..... 79

**Figure III-4** The two-dimensional INDSCAL solution obtained for 5 listeners .. 80

**Figure III-5** Subject weights for dimensions 1 and 2 ..... 81

**Figure III-6** The three-dimensional INDSCAl solution ..... 81

**Figure III-7** Badness-of-fit curve representing the fit error by nonmetric multidimensional scaling analyses. ..... 82

**Figure III-8** The two-dimensional nonmetric solution ..... 83

**Figure III-9** The three-dimensional solution of the nonmetric analyses ..... 83

**Figure III-10** The four-dimensional nonmetric solution ..... 85

**Figure III-11** (a) An example of natural fricatives without the transition part .... 88

**Figure III-11** (b) An example of LPC encoded fricatives ..... 88

**Figure III-12** The one-dimensional nonmetric solution ..... 89

**Figure III-13** The two-dimensional nonmetric solution obtained from the stimulus set 1, the cut-out fricatives ..... 90

**Figure III-14** The two-dimensional nonmetric solution of the stimulus set 2, the LPC synthesised fricatives ..... 91

**Figure III-15** The three-dimensional nonmetric solution of the LPC synthesised fricatives. ..... 92

**Figure III-16** A plot of auditory filter bank; 32 channel 1/3 octave filters ..... 95

**Figure III-17** Illustration of a time-aligned path between two segments that differ in time scale. ..... 96

**Figure III-18** Acoustic maps for the natural fricative (cut-out) segments. ..... 101

**Figure III-19** Acoustic maps for the LPC synthesised fricatives ..... 102

**Figure III-20** Two hypothetic variable sets; (a) dependent set (b) predictor set .. 103

**Figure III-20** (c) A plot of canonical scores (new set of coordinates) ..... 104

**Figure III-21** Two dimensional grid of F1 against F2 (the design of stimuli) .... 109

**Figure III-22** Subject weights for dimensions 1 and 2 ..... 110

**Figure III-23** Plots of canonical scores for perceptual and (a) Euclidean (b) Slope (c) N2D spaces ..... 112

**Figure IV-1** Examples of stimuli sets ..... 119

**Figure IV-2** The stimulus configurations for the whole syllable set from (a) interval and (b) ordinal level solutions ..... 125

**Figure IV-2** (c) The stimulus configurations of dimension 3 against dimension 1 of the ordinal level solution ..... 126

**Figure IV-3** Dimension coefficients of the stimulus set showing the subjects' weight on each dimension, for (a) interval and (b) ordinal solutions ..... 127

**Figure IV-4** Stimulus spaces from interval level analysis of whole syllable set (a) for Group A (b) GroupB ..... 129

**Figure IV-5** Stimulus spaces from ordinal level analysis of the whole syllable set (a) for Group A (b) for Group B ..... 131

**Figure IV-6** The stimulus configurations from (a) interval (b) ordinal level solutions for the no-transition set ..... 132

**Figure IV-7** Dimension coefficients of the stimulus set showing the subjects' weights on each dimensions, for (a) interval, and (b) ordinal level solutions 133

**Figure IV-8** Stimulus spaces from interval level analyses of the no-transition set (a) for Group A (b) for Group B ..... 135

**Figure IV-9** Stimulus spaces from ordinal level analyses of the no-transition set for (a) Group A (b) Group B ..... 136

**Figure IV-10** The group stimulus spaces from interval level analysis for the cut-out set ..... 137

**Figure IV-11** Subject space for interval level analysis for Group A only ..... 139

**Figure IV-12** Stimulus spaces from the interval level analysis of the cut-out set for (a) Group A, and (b) Group B ..... 139

**Figure IV-13** The two-dimensional interval level solution for the LPC22 set ... 140

**Figure IV-14** Group stimulus spaces from interval level analysis of the LPC22 set for (a) Group A, and (b) Group B. ..... 141

**Figure IV-15** Group stimulus spaces for (a) subjects 1-5, and (b) subjects 6-10 . 142

**Figure IV-16** Group stimulus spaces for (a) the first half of Group B, and (b) the second half of Group B ..... 143

**Figure IV-17** The two-dimensional interval level solution for the LPC10a set... 144

**Figure IV-18** Stimulus spaces from interval level analysis of the LPC10a set for (a) Group A, and (b) for Group B. ..... 145

**Figure IV-19** Stimulus spaces from the interval level analysis for (a) subjects 1-5, and (b) subjects 6-10 ..... 146

**Figure IV-20** Stimulus spaces from the interval level analysis for (a) subjects 11-15, and subjects 16-20 ..... 146

**Figure IV-21** Stimulus spaces from the interval level analysis of the LPC10 set for (a) Group A, and for Group B ..... 148

**Figure IV-22** Stimulus spaces from the interval level analysis of the LPC10 set for Group A (a) subjects 1-5, and (b) subjects 6-10 ..... 149

**Figure IV-23** Stimulus spaces from the interval level analysis for Group B (a) subjects 11-15, and (b) subjects 16-20 ..... 149

**Figure IV-24** Stimulus psaces from the interva level anlysis of the LPC4 set for (a) Group A, and Group B ..... 150

**Figure IV-25** Stimulus spaces from the interval level analysis for Group A (a) subjects 1-5, and (b) subjects 6-10 ..... 151

**Figure IV-26** Stimulus spaces from the interval level analysis for Group B (a) subjects 11-15, and (b) subjects 16-20 ..... 151

**Figure V-1** Auditory spaces based on Euclidean distance metric. ..... 157

**Figure V-2** 1/3-octave auditory spectra taken as an average over the whole length of the fricatives ..... 160

**Figure V-3** Plots for canonical scores, comparing perceptual and Euclidean auditory spaces for (a) whole, (b) no-transition, and (c) cut-out sets ..... 166

**Figure V-4** Plots for canonical scores, comparing perceptual and Euclidean auditory spaces for (a) Group A, and (b) Group B, for the LPC22 set .... 168

**Figure V-5** Plots for canonical scores, comparing perceptual and Euclidean auditory spaces for (a) Group A, and Group B, for the LPC10a set ..... 168

**Figure V-6** Plots for canonical scores, comparing perceptual and Euclidean auditory spaces for (a) Group A, and (b) Group B, for the LPC10 set .... 169

**Figure V-7** Plots for canonical scores, comparing perceptual and Euclidean auditory spaces for (a) Group A, and Group B, for the LPC4 set ..... 169

**Figure V-8** Auditory spaces based on the Slope distance metric. ..... 171

**Figure V-9** Auditory spaces based on the N2D distance metric. ..... 175

**Figure V-10** Five vowel auditory spaces from (a) Euclidean, (b) Slope, and (c) N2D metrics ..... 179

**Figure VI-1** 1/3 octave auditory spectra of 10 productions by 5 speakers, taken as an average over the whole length of the fricatives ..... 185

**Figure VI-2** 1/3 octave auditory spectra of 10 productions by 5 speakers, taken as an average over the whole length of the fricatives, after subtracting the mean in dB ..... 188

**Figure VI-3** 1/3 octave auditory spectra, taken as an average over 10 productions by 5 speakers ..... 192

**Figure VI-4** Canonical correlation alignments between two productions for (a) speaker 1, (b) speaker 2, (c) speaker 3, (d) speaker 4, and (e) speaker 5 .. 195

**Figure VI-5** Canonical correlation alignments between two productions for the different speakers, after RMS level normalisation ..... 197

**Figure VI-6** General auditory map of English fricatives based on 10 productions by 5 speakers ..... 198

**Figure VI-7** An average auditory map ..... 199

**Figure VI-8** The average spectrum of each fricative is placed on the corresponding region of each fricative on the auditory axes ..... 200

# *List of tables*


**Table I-1** French plosives, fricatives and nasals arranged according to place of articulation ..... 23

**Table I-2** Airline distances between ten U.S. cities ..... 27

**Table I-3** Distance scores matrix for a subject ..... 29

**Table II-1** Comparison of psychophysical and phonetic distance judgments ... 29

**Table III-1** Similarities matrix for a subject who rated the fricative syllables ... 75

**Table III-2** Symmetrised matrix of Table III-1 ..... 75

**Table III-3** Dimension weights for the three-dimensional model ..... 84

**Table III-4** Subject weights for the two-dimensional solution ..... 90

**Table III-5** A summary of previous distance metrics studies ..... 99

**Table III-6** Coordinates for the dependent (x1, y1) and predictor (x2 y2) sets, with canonical variables (new x1, y1, x2, y2) ..... 104

**Table III-7** The canonical correlation values and the probability levels for the null hypotheses ..... 105

**Table III-8** Canonical correlations between the perceptual and acoustic spaces for the LPC synthesised fricatives, based on the three different distance metrics ..... 106

**Table III-9** Canonical correlations between the perceptual and Euclidean, Slope and N2D auditory spaces of noises ..... 111

**Table IV-1** Badness -of-fit values, which show the fit error by interval and ordinal level proc-ALSCAL analyses, with the number of dimensions .... 124

**Table IV-2** Comparison of fricative ordering in the sibilance dimensions of interval and ordinal level analyses ..... 125

**Table IV-3** Comparison of average subject weights in Groups A and B respect to both interval and ordinal level analyses for each perceptual dimension .. 128

**Table IV-4** Comparison of individual subject weights in interval and ordinal level analyses for Subjects 3 and 19, for each perceptual dimension .... 128

**Table IV-5** Comparison of average subject weights in Groups A and B for each perceptual dimension (the whole syllable set) ..... 130

**Table IV-6** Badness-of-fit values for the no-transition set ..... 132

**Table IV-7** Comparison of separate average subject weights in Groups A and B for each perceptual dimension (the no-transition set) ..... 134

**Table IV-8** Comparison of average subject weights in Groups A and B for each perceptual dimension ..... 134

**Table IV-9** Comparison of average subject weights in Groups A and B for each perceptual dimensions (the cut-out set) ..... 138

**Table IV-10** A summary of phonetic interpretability of the MDS dimensions .. 152

**Table IV-11** (a) Correlations between the whole syllable, no-transition, and cut-out sets of the place dimension in their group stimulus spaces ..... 153

**Table IV-11** (b) Correlations between the whole syllable set and the other sets with respect to the place dimension ..... 153

**Table V-1** Canonical correlations between the perceptual and auditory spaces for each stimulus set ..... 163

**Table V-2** Canonical correlations between the perceptual and auditory spaces for each stimulus set, based on the Slope distance metric ..... 173

**Table V-3** Canonical correlations between the perceptual and auditory spaces for each stimulus set, based on the N2D distance metric ..... 177

**Table VI-1** The mean duration, in ms, of the fricative portions of two repetitions from the five different speakers ..... 183

**Table VI-2** The mean RMS levels of fricatives across different productions, measured from the whole length of the fricative sections ..... 184

**Table VI-3** Canonical correlation coefficients between the two different productions of each speaker ..... 194

**Table VI-4** Canonical correlation coefficients between the two separate productions of the fricatives by the same speakers, after loudness normalisation 196

# *Chapter I. Introduction*

## 1 Invariance and variability in speech perception

This study is about speech perception and its relations with other recognised levels of speech communication such as acoustics, phonetics, and articulation. By speech communication we mean how the speaker converts a message into speech sounds, which are then translated into a meaningful message for a hearer. The movement of the vocal organs of the speaker produces sound waveforms which cause air pressure change that propagate from speaker to hearer. When the ear responds to pressure changes, the peripheral auditory system generates nerve impulses. These auditory nerve impulses travel through the auditory nerves to the brain. Linguistic processes in the brain are stimulated and the speaker's intended message is decoded in the hearer's brain. This overall view of communication is often called the *speech chain* (Denes & Pinson, 1993) and the various different levels of speech communication are identified according to it.

By speech perception, we refer to the decoding process from which acoustic signals are converted into messages. From this simple description of speech perception, several domains of processing can be observed. When the speech waves activate the peripheral auditory system, it is in the acoustic domain. When the acoustic waveforms cause the nerve impulses to fire and carry information to the brain, it is operating in the physiological domain. In the brain, these impulses are incorporated into linguistic processing, which brings about recognition of the speaker's message. Thus the process has reached the linguistic domain.

Although it is obvious that the speech signal needs to change from one medium to another, how these changes are carried out is not clear. So far we have not found any units of the speech signal which can be directly matched to units of the linguistic message. Rather, the units of speech signal have to be transformed into a structure[1] suitable for

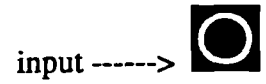
[1]The precise nature of such a structure, whether it be acoustic cues, feature specifications, segments or syllables, is itself controversial. For example, in the TRACE model (McClelland & Elman, 1986), this basic linguistic unit is the feature; in the COHORT model this unit is the phoneme (Marslen-Wilson & Tyler, 1980).

lexical access and for reference to other previously stored knowledge. The problem of defining such a linguistically relevant unit — let's call it a *phonetic segment* — lies primarily in the nature of the acoustic signal. For example, the exact form of a speech sound depends not only on the phonetic segment intended by the speaker, but also the influence of neighbouring sounds (coarticulation) and of its structural position (melodic constraints). Also, the acoustic nature of the speech signal varies vastly, depending on the environment; whether it is heard inside a moving train or a quiet room; whether it is articulated by an adult or a child; whether it is articulated carefully or hurriedly. All these different conditions contribute to the variability of speech signals at the acoustic level. As a result, there are few cases of a one-to-one mapping between an acoustic feature of the signal and a phonetic feature of the message.

Despite this lack of invariance in the speech signal, the linguistic identity of a phonetic segment is very robust. Somehow the hearer is able to extract the relevant acoustic cues from the speech signal which provide the necessary information to reconstruct the phonetic segment originally intended by the speaker. Therefore, the central problem of speech perception can be summed up as the correspondence between the *constant phonetic percept* that encodes the pronunciation of speaker's message and the *variable acoustic signals*.

## 2 Methods of investigation

The problem of investigating the intermediate process between acoustic signals and phonetic percept is a formidable one, since this process is carried out below the level of consciousness. As it is schematised below, the only concrete manifestations of this process are the input (signals) and output (percepts).




How can we determine the nature of this 'black box'? Somehow we must find inputs to which the black box will react and produce different outputs. Early studies have therefore been concentrated on identifying perceptually important aspects of speech signals, by studying spectrograms and by the careful manipulation of synthetic speech. A common

practice in speech perception experiments is for the experimenter to control the perceptually salient part of signals or listening conditions in the laboratory and ask subjects to identify what they hear under these conditions. In this way, the experimenter can be sure of exactly what the input description is, and a set of hypotheses about the perceptual significance of certain acoustic characteristics can be tested.

In this way it has been found that, for example, the formant pattern is important for vowel perception, release burst for stops, high frequency spectral pattern for fricatives, and so on. Specifying details of perceptual cues beyond these rough descriptions is, however, more problematic. For example, it was found that contextual effects cause the specific cues for each segment to interact. Depending on which other segments are adjacent, the cues for each segment are obscured or dispersed over several segments. Conversely, any particular time-window may contain cues for several segments. In general, findings concerning the perceptual importance of individual cues have led us no closer to finding acoustic invariances corresponding to phonetic representations.

Because this description of the speech input based on the fine details of acoustic cues is so difficult to relate to any obvious phonetic units, it has given rise to alternative descriptions of speech signal processing — some based on articulatory inferencing (e.g. Motor theory of speech perception, Liberman & Mattingly, 1985), others centred on the invariant aspects of the signal (Acoustic invariance theory, Stevens & Blumstein, 1979), for example.

In Motor theory, instead of addressing the problem of acoustic variance, proponents have suggested invariant articulatory gestures as a basis for decoding the message. It is claimed that, somehow, the listener is capable of extracting the speaker's neuron motor commands from the acoustic signals. Just how this process is achieved is never made clear, however.

Acoustic invariance theory suggests that it is an inappropriate perspective on the speech signal, focusing on steady states connected by transitions, which makes speech appear more variable than it is. Instead, if the focus is on the fine acoustic detail in the periods of rapid change between steady states (eg. at the release burst for stops), the invariant properties of the signal can be discovered. However, the demonstration of these kinds of 'absolute' invariants has been controversial; furthermore, its success has not been

sufficiently general.

These kinds of finely detailed acoustic signal studies, whether to determine perceptually salient acoustic cues or invariants, have failed to resolve the issue of 'acoustic variability versus constant phonetic percept'. This seems to indicate a weakness in this experimental paradigm. The ultimate aim, after all, is not to predict listeners' responses to every possible stimulus, but to build models of phonetic processing in general. The weakness of an approach only based on minimal contrasts is that it misses the simple and robust features of sounds which define the major categories of syllable, voice, manner or place. The reductionist approach may also give rise to models that are particularly speaker- listener- or language-dependent. The response of listeners to particular category labels can be strongly influenced by their previous language exposure (e.g. Simon & Fourcin, 1978). This suggests a more 'holistic' approach to speech perception; perhaps we do not need to identify the specific processing associated with each phonetic category, but instead, see categories as existing in a region of perceptually acceptable realisations in a phonetic space of a small number of dimensions.

## 3 Study of spatial representations

### 3.0 Introduction

In this section we will be introducing the notion of *space* and how it can be applied to a more holistic approach to research within various areas of speech perception: to articulation, phonetics, acoustics, audition and perception. We highlight a certain asymmetry in the studies of spatial representation such that, on the one hand, we find concreteness and explicitness in the notion of space concerning vowel studies, and on the other hand, abstractness and ambiguity in studies of consonants. If, however, we suppose that consonants and vowels are subject to the same physiological and linguistic constraints, this asymmetry in their spatial representations is somewhat anomalous. Or is it possible that consonant inventories are structured according to principles that are different from those which apply to vowel systems? Before addressing such issues we will be examining the importance of spatial formulations in studies of speech processing.

### 3.1 Articulatory/Phonetic space

At the beginning of this century, Daniel Jones had managed to condense all the possible phonetic representations of vowels onto a two-dimensional graph, which is known as the *'vowel quadrilateral'*. Figure I-1 shows a variant of Jones' vowel quadrilateral (see Jones, 1975). The horizontal dimension of the chart represents tongue advancement whilst the vertical dimension is related to the tongue height. The four corners and the intersections of the boundary and horizontal lines of the chart correspond to Jones' eight cardinal vowels; these points signify the human articulatory boundaries (limits) for vowel articulation.

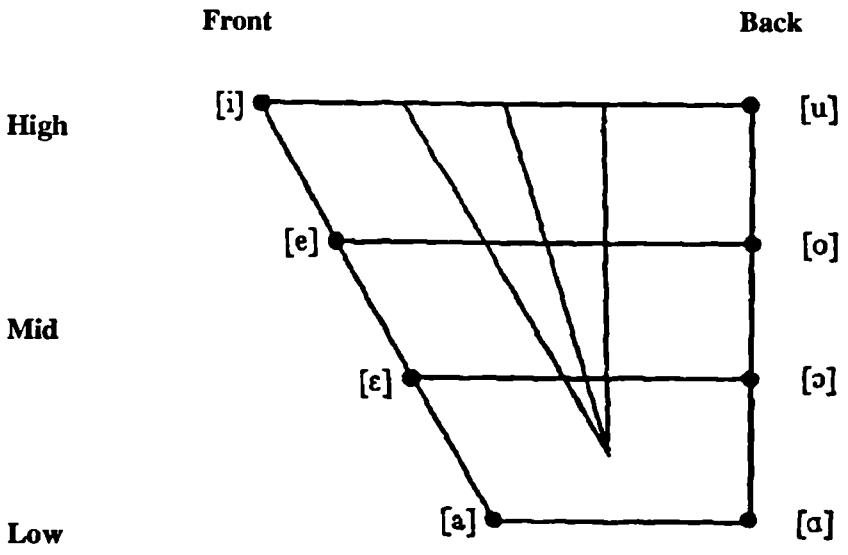



**Figure I-1.** The vowel quadrilateral; after Gimson (1989).

The vowels of any of the world's languages can be represented by their positions in the vowel quadrilateral. For example, a three-vowel language like Warlpiri, spoken in central Australia, can be depicted in terms of a triangular vowel space:

| | **Front** | **Back** |
|---|---|---|
| **High** | i | u |
| **Low** | a | |

Five-vowel languages like Spanish, Russian and Japanese employ a pentagonal vowel space:

| | Front | | Back |
|---|---|---|---|
| **High** | i | | u |
| **Mid** | e | | o |
| **Low** | | a | |

A common pattern of distribution in vowels is that they tend to be maximally dispersed in the vowel space, in order to maintain perceptual distinctness, yet sufficiently compact for ease of articulatory manoeuvre. Thus we may predict the configurations of a seven vowel space like Italian to be as follows:

| | Front | | Back |
|---|---|---|---|
| **High** | i | | u |
| **High mid** | e | | o |
| **Low mid** | ɛ | | ɔ |
| **Low** | | a | |

This basic symmetry and patterns of distribution are generally maintained throughout the world's languages, whatever the size of the inventory. These observations could serve as the basis for the formulation of universal principles governing vowel systems. For example, Ohala (1979) concluded that a principle of maximal phonetic differences[2] adequately predicts the placement of vowels in the available phonetic space. Therefore, the spatial representation of vowels plays a crucial role in understanding and formulating the phonological system.

Nothing as elaborate as the quadrilateral has been proposed for consonants but some symmetries and patterns have been observed. Clark & Yallop (1990) illustrate this point with French. Table I-1 is a reproduction of the chart in Clark & Yallop (p135).

| | Labial | Dental/ alveolar | Back |
|---|---|---|---|
| voiceless plosives | p | t | k |
| voiced plosives | b | d | g |
| voiceless fricatives | f | s | ʃ |
| voiced fricatives | v | z | ʒ |
| nasals | m | n | ɲ |

**Table I-1.** French plosives, fricatives and nasals arranged according to place of articulation; after Clark & Yallop (1990).

[2]Here phonetic differences incorporate both articulatory and auditory differences.

Indeed many languages of the world seem to maintain this 'place of articulation' and 'voicing' distinction, albeit with further subdivisions (See Ladefoged & Maddieson, 1996).

Another universal pattern of consonant distribution is studied by Lindblom & Maddieson (1988), who examined consonant data from over 300 languages using the UCLA Phonological Segment Inventory Database (UPSID). They found that, although the number of consonants in a language may be as few as six, and therefore any given inventory could, in theory, consist of only sonorants or only obstruents, in practice languages tend to have about 70% obstruents and 30% sonorants. Because this proportion is independent of the size of an inventory, they suggest that this reflects a "phonetic universal" which is described as "the physical characteristics of the regions of the 'phonetic space' that obstruents and sonorant consonants range in" (p66).

Incorporating motor and perceptual effects into the characterisation of consonantal distribution, they have formulated a universal principle as follows:

> Consonant inventories tend to evolve so as to achieve maximal perceptual distinctiveness at minimum articulatory cost. (p72)

This principle of balance between two major components of a phonological system was schematised in Figure I-2.

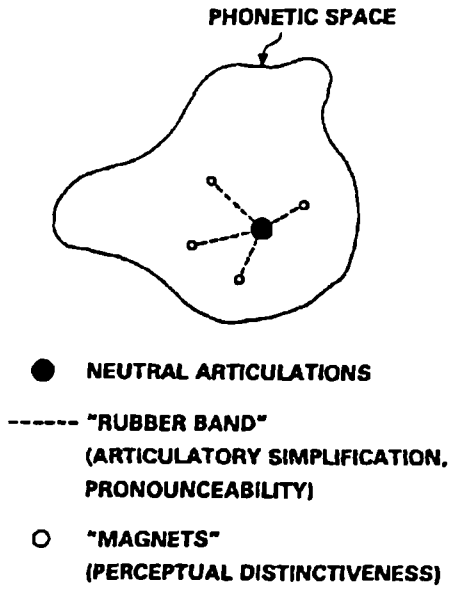



**Figure I-2.** Phonetic space' for consonant systems of the world languages; after Lindblom & Maddieson (1988)

We shall not delve into the details of the analogy between the diagram and their universal 'consonant-system', nor shall we offer any critique of such a universal principle. What needs to be pointed out from the figure is that the concept of 'space' is also central to an understanding of the consonant system, and that this notion is probably common to the processing of both vowels and consonants.

What distinguishes the phonetic space of consonants from that of the vowel space, however, is the fact that the limits of articulation cannot be defined for consonants, and instead, the space adopts a flexible form which can be expanded or shrunk depending on a phonological inventory. Thus the concept of space has been rather abstract for consonants, and as a consequence, spatial representations in acoustic and auditory domains have not been formulated. On the other hand, the notion of space has been strengthened and, has become indispensable in describing the physical properties of vowels, and their relationship with phonetic/articulatory properties. This point will be illustrated in subsequent sections.

### 3.2 Acoustic/auditory space

The notion of *acoustic space* for vowels is suggested as far back as in 1948, by Joos. He pointed out the possibility of representing vowels on F2/F1 plane in order to draw a parallel between the properties of vowel production described in Jones' quadrilateral and vowel acoustics. He had realised that the first formant can be related to tongue height and the second formant to tongue advancement. This means that if we take the measurements of the first two formants of the vowels of a language and place them on F2/F1 plane the result will resemble the articulatory chart for the vowels. This was only an approximation, however, and he had found that these formants need to be plotted on a nonlinear (logarithmic) scale, "like the musical scale" (p52). Later studies of the properties of the auditory periphery show that nonlinear transformations of the frequencies were necessary to model peripheral auditory processes. Therefore, what is known as F2/F1 chart of vowels is, in fact, something which describes the *auditory space* of vowels.

One of the earliest studies of the auditory space of rounded and unrounded cardinal vowels was given by Peterson (1951), based on the sustained unrounded and rounded vowels spoken by a phonetician (native speaker of American English). The

frequencies have been arranged in accordance with the so-called Mel scale (Stevens and Volkmann, 1940). This scale represents perceptually equal intervals of pitch as equal distances along the frequency scale. This is shown in Figure I-3.

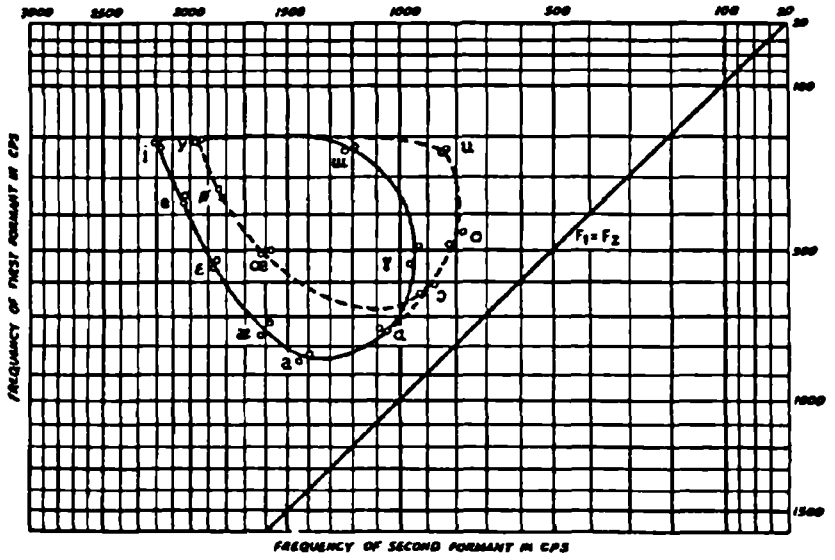



**Figure I-3.** A plot of Mel(F2) against Mel(F1) of cardinal vowels produced by a male phonetician; after Peterson (1951).

The points for the unrounded vowels are connected by a solid line; a dashed line connects the rounded vowels. We can observe an excellent resemblance to the articulatory/phonetic map in Figure I-1.

More recently, it has been suggested that the early stages of the auditory system can be treated as a bank of bandpass filters. Numerous studies have been carried out to estimate the auditory filter bandwidth and centre frequency of each filter. It was found that the filter bandwidth increases with filter centre frequency in a nonlinear fashion, and the scale is given the cumulative number of filters as a function of frequency. Examples of such scales are the Bark or z scale (Zwicker *et al.*, 1979) and ERB-rate scales (Glasberg & Moore, 1990). An example of auditory vowel space based on Bark transforms is by Sydral & Gopal (1986). One axis is (z(F1)-z(F0). The other axis is (z(F3)-z(F2)).

Therefore, we have seen that although the fine details of the auditory space for vowels needed adjustment, it clearly illustrates that Joos' (1948) early intuitions were correct, in that some sort of simple relationship between phonetic/articulatory and

acoustic/auditory space exists in vowels, and that this relationship has been made explicit by comparison of spatial relations in the two different domains.

### 3.3 Perceptual spaces

We have examined phonetic space, which was based on articulatory manoeuvres, and this space was closely related to the acoustic/auditory space, which was essentially the F1/F2 plane. So we have found a close relationship between the phonetics, the articulation and the acoustics/audition of vowels. The question remains as to how these spaces bear on the structure of the *perceptual space* of vowels. Before we can answer this, we must explain the notion of vowel perceptual space and how this can be examined.

The nature of perceptual space is quite different from the other two types of space we have examined so far, in that we do not know the axes of this space, nor are we going to predetermine the nature of its axes as we have done in the two previous cases - tongue position and formant values. Instead our aim is to let the axes emerge out of a perceptual study of similarities. This sounds all very mysterious and abstract. An analogy may help to clarify this point.

Let us pretend that we have no knowledge of the map of the United States and we only know the distances between the American cities as below:

| CITIES | ATLA. | CHIC. | DENV. | HOUS. | L.A. | MIAMI | N.Y. | S.F. | SEAT. | WASH D.C. |
|---|---|---|---|---|---|---|---|---|---|---|
| ATLANTA | | 587 | 1212 | 701 | 1936 | 604 | 748 | 2139 | 2182 | 543 |
| CHICAGO | 587 | | 920 | 940 | 1745 | 1188 | 713 | 1858 | 1737 | 597 |
| DENVER | 1212 | 920 | | 879 | 831 | 1726 | 1631 | 949 | 1021 | 1494 |
| HOUSTON | 701 | 940 | 879 | | 1374 | 968 | 1420 | 1645 | 1891 | 1220 |
| LOS ANGELES | 1936 | 1745 | 831 | 1374 | | 2339 | 2451 | 347 | 959 | 2300 |
| MIAMI | 604 | 1188 | 1726 | 968 | 2339 | | 1092 | 2594 | 2734 | 923 |
| NEWYORK | 748 | 713 | 1631 | 1420 | 2451 | 1092 | | 2571 | 2408 | 205 |
| SANFRANCISCO | 2139 | 1858 | 949 | 1645 | 347 | 2594 | 2571 | | 678 | 2442 |
| SEATTLE | 2182 | 1737 | 1021 | 1891 | 959 | 2734 | 2408 | 678 | | 2329 |
| WASHINGTON D.C. | 543 | 597 | 1494 | 1220 | 2300 | 923 | 205 | 2442 | 2329 | |

**Table I-2.** Airline distances between ten U.S. cities; after Kruskal & Wish (1978).

How can we produce a map of the United States from this table of distances between the cities? This problem can be solved by applying Multidimensional scaling (MDS) analysis. This is a mathematical technique which allows us to represent objects as a set of points in a space where the dimensionality of the space and the positions of points reflect the distance between the objects. The outcome of MDS on the distances of U.S. cities is as below:

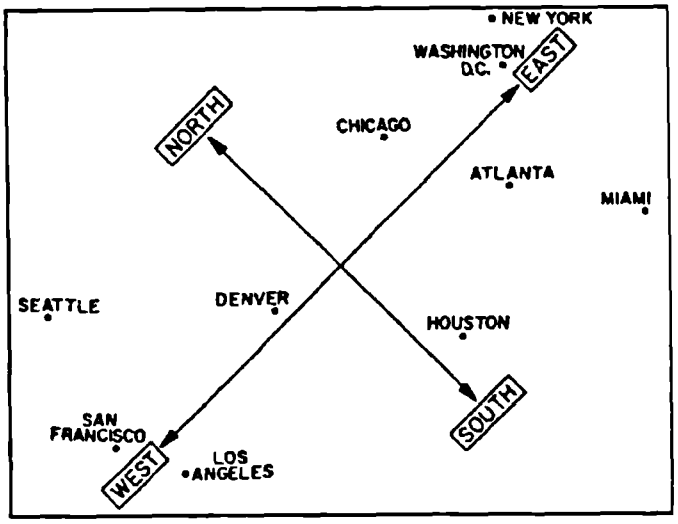



**Figure I-4.** Multidimensional scaling map of the airline distances among ten U.S. cities; after Kruskal & Wish (1978).

We can easily recognize this to be the geographical locations of ten U.S. cities, save for the orientation of the map.

The problem of obtaining a vowel perception map can be solved if we can find out about the *perceptual distances* between the vowels. How could we measure this perceptual distance? In other words, how could we *quantify* the perceived differences between the sounds? The relative distance scores can be obtained by asking subjects to rate (dis)similarity amongst a particular set of sounds. These scores can be represented as a (dis)similarity matrix of stimuli against response, as in the distances between the cities.

A good example of this kind of work is by Rakerd (1984). He obtained the distances between 10 American English vowels /i ɪ ɛ æ ɑ ɔ ʌ o ʊ u/ in a /dVd/ context by means of a triadic comparison test. He asked subjects to choose the most similar pair

and the most dissimilar pair among three sounds they had heard. A distance score of 1 and -1 are assigned to each respectively. An example of accumulated distance score is given below:

| | i | ɪ | ɛ | æ | ʌ | ɑ | ɔ | o | ʊ | u |
|---|---|---|---|---|---|---|---|---|---|---|
| i | | | | | | | | | | |
| ɪ | 6 | | | | | | | | | |
| ɛ | -2 | 2 | | | | | | | | |
| æ | -4 | -4 | 5 | | | | | | | |
| ʌ | -2 | -4 | 7 | 1 | | | | | | |
| ɑ | -8 | -4 | 1 | 4 | 1 | | | | | |
| ɔ | -5 | -3 | 0 | 1 | 4 | 5 | | | | |
| o | -6 | -5 | -2 | 0 | 2 | -3 | 3 | | | |
| ʊ | -5 | 1 | 4 | -1 | 5 | -2 | 1 | 3 | | |
| u | -3 | -2 | 2 | -1 | 1 | -1 | -4 | 4 | 8 | |

**Table I-3.** Distance scores matrix for a subject who rated vowels in consonantal context; after Rakerd (1984)

23 subjects, therefore 23 such matrices, have been obtained. Since there is more than one distance matrix to the input of the MDS procedure, the differences between individual subjects need to be reflected in the spatial solution. This can be done by applying Individual differences scaling (INDSCAL). This technique not only calculates the relative locations of the objects in a space, but also calculates the relative weights that each subject places on a particular dimension in order to find an optimal orientation of the space[3].

By applying an INDSCAL analysis, Rakerd has obtained a three dimensional solution for the vowels, of which the first two dimensional space is shown as follows:

[3]Details will be explained in §III.2.4

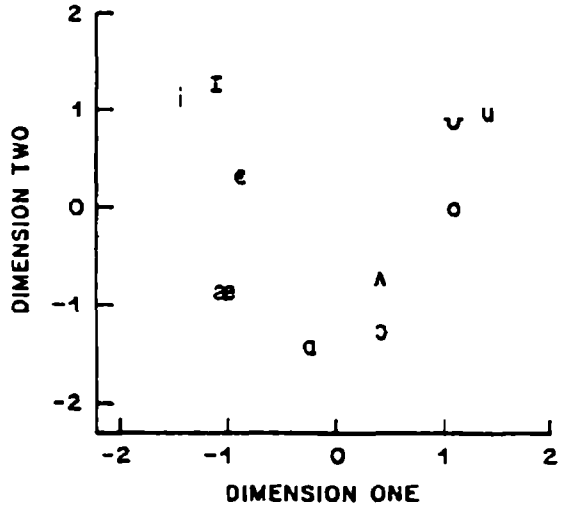



**Figure I-5.** First two dimensions of MDS solution of the vowels; Rakerd (1984)

The dimensions clearly correlate with the tongue advancement (F2) and height (F1) distinctions that we examined in the previous two sections. Other MDS studies on vowels also showed that the first two dimensions are stable and always correlate with those properties of vowels (See Fox, 1983; Rakerd & Verbrugge, 1985).

In sum, the MDS studies show that the concept of perceptual distances between sounds is crucial in investigating a set of sounds without *a priori* assumptions about what underlying cues or articulatory manoeuvres are necessary. Transformation to spatial representations from the distance scores is important, since the large number of stimuli can be condensed into a low dimensional space. It was shown that the principal dimensions that came out of the spatial analyses were closely related to spatial representations of other domains of speech processing.

At this point, we have discussed all the possible spatial representations concerning vowel perception — the phonetic/acoustic, acoustic/auditory, and perceptual spaces — and we have learnt that all of these spaces have the same basic organisation. The implications for such spatial relationship for vowels will be discussed in the next section.

### 3.4 Spatial representations and modelling of perceptual processes

We have seen that the two principal dimensions of the phonetic space of vowels are defined by tongue height and advancement. The auditory space of vowels based on F1/F2 is closely related to this phonetic/articulatory space. The perceptual reality of these dimensions is endorsed by MDS analyses of perceived (dis)similarity between the vowels.

The implication of this simple and clear correspondence between these different vowel spaces is that there must be a direct relationship between all these different domains of processing. That is, the perception of speech sounds need not be treated specially. We have mentioned in §1 that the issue in studies of the speech processing mechanism is to find out what is contained in the intermediate process between input (acoustic signal) and the output (the percept). From the studies of vowel spaces it was shown that their organisation in perceptual space could be adequately explained by observing the organisations in the auditory/acoustic space. This means that what we had thought of as a 'black box' (in §2) is no longer beyond the scope of our knowledge. After the nonlinear auditory frequency and loudness transforms, the auditory patterns can be directly matched to corresponding phonetic percepts, which are ready to be used for higher level speech processing.

A recent spatial model of vowel perception, put forward by Rosner & Pickering (1994; p117), is based precisely on such an assumption. They suggest that the vowel plane contains an auditory prototype for each vowel in the language. The identification rate for each vowel is inversely related to the distance between a current point and its nearest prototype on this plane. This results in an acceptable auditory area which can predict the correct identification of a particular vowel on the auditory map. The auditory plane contains various nonlinear frequency and loudness transformations to model the auditory peripheral processing.

One problem in suggesting this kind of simple and direct 'auditory model' (as opposed to the Motor theory, for example) of speech processing, however, is that the phonetic/articulatory and acoustic spaces of vowels are so closely related, that it is hard to measure either any phonetic influence there might have been, or any kind of articulatory referencing in the perceptual judgements of the vowels. In other words, this kind of direct relationship was anticipated because of inherent similarities in the phonetic /articulatory

and the acoustic forms of vowels.

But this statement is only true for vowels. No language in the world comprises only vowels, however. Since we have only accounted for one half of the system, it is impossible to make any language-wide generalisations. Categorical speech perception studies suggests that vowels are continuously perceived (indefinite number of classes), while consonants are perceived categorically (fixed set of classes)[4]. Thus, it would not be acceptable to infer perceptual processes for consonants from vowel results. On the other hand, as already stated at the beginning of this section, both vowels and consonants are likely to be processed within the same physiological and linguistic constraints. If this is so, both vowel and consonant sounds need to be studied using the same approach. So far in speech perception research, it is only the differences between the two which have been emphasised. In this study, we are adopting essentially the same approach to the two types of speech sounds, in an attempt to give a more generalised view of speech processing.

## 4 Objectives

The objective of this research is to extend the studies of spatial representations on vowels to a certain set of consonants. Fricatives and fricative-like sounds are chosen since they provide a convenient medium in which the knowledge of vowel analyses can be applied to consonant studies. A set of questions needs to be postulated here:

i) Would we find the same correlation between auditory and perceptual spaces for fricatives, as we have already found to be the case in vowels? If so, this relationship shows that the basis for human perception is shared between consonants and vowels alike.

ii) Alternatively, if the relationship is not straightforward, what sorts of results are expected? Would we be able to observe phonological or phonetic influences in the processing? Fricative data may be able to clarify the issue, since the phonetic properties of fricatives are not so simply linked to their acoustic properties. Thus we should be able to separate the influence of phonetics from acoustics in the

[4]It is widely believed that vowels are not categorically perceived. Different views are put forward in later categorical studies. See Rosen & Howell (1987).

process of speech perception. It may be the case that spatial representations from the MDS process will show some distinct phonetic groupings, which cannot be explained simply by referring to their auditory organisation.

iii) In addition, would the specifications of auditory transforms used in vowels be true for some consonants?

By answering these questions we will be able to address the issue of whether vowels and consonants are processed in the same way, and give a better overview of the speech processing mechanism.

## 5 Outline of thesis

In this chapter, we have introduced the general notion of spatial representations, which is fundamental to the investigation of the relationship between various different domains of speech perception. In addition, we have shown how this notion can be applied to the general mechanism of speech processing. Chapter II presents a detailed review of vowel studies which are concerned with the relationship between perceptual and physical spaces. The main perceptual and acoustic analyses of the present study will be modelled on these studies. Some of the attempts made in this direction of research for consonant data are also briefly reviewed. The second section is concerned with auditory distance modelling studies of speech perception. We will review such classic works as Klatt (1982a), together with subsequent models and distance metrics on vowels. The last section examines previous acoustic and perceptual studies carried out on fricatives and discusses the viability of applying the techniques described in the previous two sections for vowels. Thus this chapter sets out all the necessary background for the main experimental work to follow.

Chapter III develops and applies to a small set of fricative sounds the techniques for extracting perceptual similarities and transforming them into spatial representations. Various distance metric models are described and given a preliminary test on perceptual data. The outcome of this preliminary experiment guides the design of the main perceptual, auditory, and production experiments in subsequent chapters.

Chapter IV is devoted to the construction and application of a large scale

perceptual experiment on fricative and fricative-like sounds. Detailed MDS analyses are carried out on the perceptual similarity data. As a result, the spatial configurations of fricative perception are established here.

Auditory modelling, on the spatial configurations obtained in Chapter IV, forms the central theme of Chapter V. Spatial representations of auditory spectra are obtained by applying the distance metrics described in Chapter III. Quantitative relations between the perceptual and auditory spaces are analysed by canonical correlation analyses.

Chapter VI investigates speaker-dependent variations in auditory modelling from multiple productions by groups of speakers. The first analysis to be carried out is concerned with the spectral aspect of the multiple speaker data, and this allows us to obtain 'average' auditory spectra. The second analysis concerns spatial aspects, and enables us to observe intra- and inter-speaker variations. Following this, the general and average auditory spaces of fricatives are drafted and their acoustic correlates are also identified.

Finally, in Chapter VII we offer a general discussion of the results of production, perception and auditory experiments, and assess the extent to which the general objectives have been achieved. We also discuss the implications and contributions of this thesis to our understanding of the topic of speech perception.

# *Chapter II. Previous studies*

## 1 Introduction

The previous chapter introduced one of the major issues in speech perception, the relationship between invariant phonetic percept and the variable acoustic signal. A brief overview of how this area has been investigated, and the advantages to be gained from addressing the issue in terms of spatial relationships were presented. Such an orientation of research has been very popular for vowel studies, and fundamental to our understanding of vowel perception. Although there is no motivation for suggesting a different mechanism for consonant perception, this method of investigation has not, however, been taken up in the same way for consonants. This is the motivation in the present thesis for applying to fricatives the spatial analyses developed on vowels, in order to broaden this view of the speech processing mechanisms.

In this chapter, we first review the key vowel studies, together with the few attempts made concerning consonants. §2 is about relevant spatial studies on vowels and consonants. §3 discusses the major auditory distance studies, from which more faithful auditory spatial configurations of speech signals may be obtained. §4 is a brief review of studies concerning articulatory, acoustic and perceptual aspects of fricatives.

## 2 Spatial relationship

### 2.0 Introduction

In §I.3, we have seen how the spatial representations of speech sounds were possible in various domains of speech processing — phonetic, auditory, and perceptual. However, these have been restricted to each individual domain and the relationship between all the different domains has merely been inferred from these individual results. For example, there have been a number of studies investigating the perceptual reality of a particular set of linguistic features (e.g. Singh & Black, 1966; Singh & Woods, 1970), but in most cases, only *post hoc* acoustic interpretations of the perceptual dimensions have been added (e.g. Fox, 1983; Rakerd & Verbrugge, 1985; Soli & Arabie, 1979). Here, however, our

primary interest is of studies which have investigated the relationship between all the different domains — phonetic/articulatory, acoustic/auditory, and perceptual.

We first examine vowel studies which have successfully demonstrated a simple relationship between all these different domains, without recourse to predetermined acoustic or linguistic parameters (in §2.1). We also examine a vowel study which was concerned with a 'limited' phonetic space and with the investigation of phonetic influence in terms of spatial representations (in §2.2). The present study will be loosely modelled on these vowel studies.

There have been no corresponding consonant studies, but spatial reinterpretations of the Miller and Nicely (1955) data are discussed below to highlight inconsistencies between vowel and consonant studies in the field (in §2.3). In addition, a recent three-dimensional spatial model of stop consonants is reviewed (Sussman, 1991).

## 2.1 Perceptual and physical spaces in vowels

### 2.1.0 Introduction

In this section, the works by Pols *et al.* (1969) and Klein *et al.* (1970) are reviewed. These two studies are very closely related; in fact, the second is a multiple-speaker version of the first experiment. Despite the early dates of these works, they address issues which remain relevant to present day speech perception studies, particularly in terms of the analysis techniques employed. The reason why these studies have not been treated as 'classics' in the speech perception field is perhaps due to their engineering-bias, i.e. for automatic speech recognition, and their lack of reference to phonetic theories or speech perception studies.

The ideas generated by reviewing these papers will be used to motivate our own experiments in later chapters.

### 2.1.1 Pols, Kamp & Plomp (1969)

The experimental paradigm (as described in following sections) applied by this work appears attractive, by virtue of the following points. However, each point will be qualified in turn.

(a) it adopts a holistic approach to the speech processing mechanism, by dealing with the Dutch vowel system as whole;

(b) it relates perceptual dimensions directly and quantitatively to the auditory dimensions;

(c) it applies an auditory analysis which is not only confined to well-defined formant analysis, but can be used to the analysis of any sound spectra;

(d) it relates auditory dimensions to the traditional phonetic dimensions of vowels.

In relation to (a), the stimuli set used was a replicated one-pitch period section taken out of each naturally produced Dutch vowel in an /hVt/ context by one of the authors. For each stimulus, the single period was resampled and reiterated to make up a 405 ms sound on a fundamental frequency of 123.5 Hz. The loudness levels of the stimuli were made equal according to the perceptual mean loudness deviation, which was determined by pairwise loudness judgements. However, we will see later that the artificial nature of these stimuli is the major weakness of this study.

The fundamental assumption in this study, which was basis for the analysis in (b), is quite radical, in that a one-to-one correspondence is implied between physical and perceptual dimensions in vowels:

> "It may be expected that the perceptual space is related to the physical space, since at least some of the physical dimensions must correspond to the way in which subjects discriminate between stimuli. The minimal number of physical dimensions required to describe the differences between vowel sounds can be considered as an indication of the number of perceptual dimensions required." (p459)

Therefore, the multidimensionality of vowels manifested in their physical properties — such as formant frequencies, amplitude levels of formants, spectral tilt, fundamental frequency, and formant bandwidths — must be reflected in their perceptual dimensions.

In order for (b) to be possible, there were three separate parts to this study — auditory spatial representation, perceptual space analysis, and correlation between auditory and perceptual representations. In the third part, where the correlation between auditory and perceptual dimensions are related, we can verify the point stated in (b).

Meanwhile, in the auditory and perceptual analyses we will verify the points (c) and (d).

With respect to (c), the way in which the auditory dimensions of the stimuli were extracted was the most original, in the sense that the issue of whether the formant values or the whole spectral shapes have perceptual significance did not have to be addressed.

A preliminary analysis of the vowel spectra was undertaken using 18-channel, 1/3 octave band pass filters, which resembles the ear's critical band pass analyses. The outputs in decibels in the 18 frequency channels produce an 11×18 data matrix. A spatial model of 11 vowel points in an 18-dimensional space will produce a perfect fit of the data. The goal of auditory analysis, however, is to determine the minimal number of dimensions required to model the data with maximal variance in the data accounted. For this purpose, Principal Component Analysis (PCA, Harman, 1967) was used. The first three factors of PCA (new auditory dimensions 1 to 3) covered 81.7% of the total variance. The three-dimensional PCA space of vowels is presented in Figure II-1.

Since this method of obtaining auditory space does not specifically refer to 'well-defined' formants, it can be applied in the analysis of any spectra, although the challenge in applying this method to fricatives would be to account for the dynamic changes in the spectra (See §III.3.2.2).

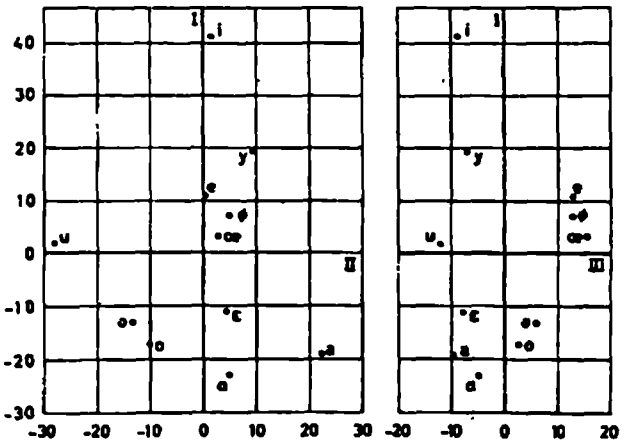



**Figure II-1.** Projections of the 11 stimulus points on two perpendicular planes of the three-dimensional physical space; after Pols *et al.* (1969).

The perceptual distances between the stimuli were extracted by a triadic comparison method from 15 listeners. The data collection process was fully interactive. From accumulated scores of similarity judgements, similarity matrices were obtained for each subject and they were averaged out to give one similarity matrix. The matrix was, in turn, converted to spatial distances, represented on suitable dimensions, by Kruskal's (1964) 2-way multidimensional scaling technique. Three dimensional solution was accepted as optimal, and this is reproduced in Figure II-2.

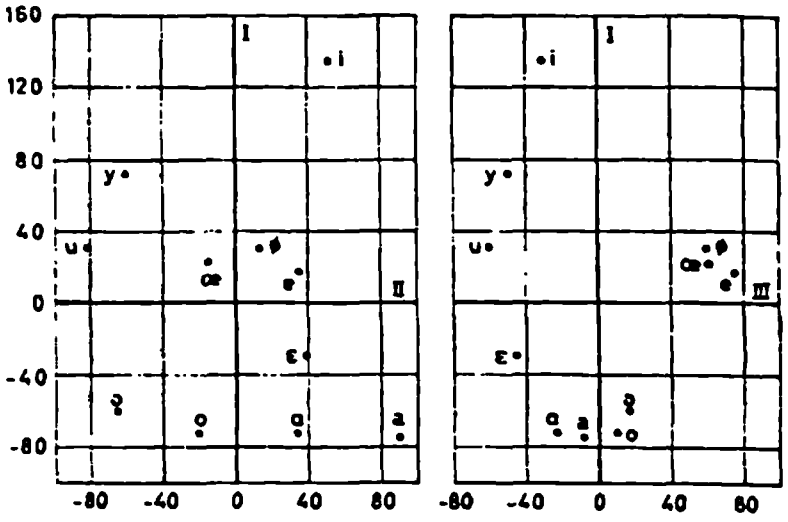



**Figure II-2.** Three-dimensional perceptual space of 11 vowel stimuli; after Pols *et al.* (1969).

In relation to point (d), they comment that the plot of the perceptual dimensions "are in accordance" with the traditional vowel chart configuration. We will shortly argue that this cannot be strictly maintained, but for the moment we will concentrate on the relationship between the physical and perceptual space.

In order to relate perceptual dimensions to the auditory dimensions, they used a 'canonical matching procedure' (See §III.3.3), which allows a comparison between two independent multidimensional spaces. The orthogonal correlations between the two spaces were very high (r = 0.992, 0.971, 0.742). They concluded that only a few physical parameters from the outputs of 1/3 octave filter analyses may be adequate for predicting human- and machine-recognition. Thus, the point given in (b) is finally verified.

Now, let us examine shortcomings present in this study. As mentioned earlier, the

major shortcoming of Pols *et al.* is unclear stimuli design. This can be illustrated by a series of self-contradictory remarks made in relation to the nature of the stimulus set. Firstly, the plot from the PCA configurations (Figure II-1) was reported to be "somewhat different" from the ones found in other experiments. This discrepancy was mainly attributed to the set of experiment materials used:

> ... the signals used do not necessarily represent the average Dutch vowel sounds by which name they are described in this article, owing to the use of only one period out of vowels spoken by only one person. (p464)

However, they have defended the legitimacy of using such simplified speech materials from the evidence they have gathered from an earlier perceptual experiment:

> This does not mean, however, that the signals were not recognisable as the appropriate vowels. Despite the modifications that were carried out on the sounds, and despite their isolated presentation, 10 of the 11 signals were practically unanimously denominated by 15 subjects as the vowels which were originally pronounced. (p464)

Also, they claim that the plot of the perceptual dimensions (Figure II-2) "are in accordance" with the traditional vowel chart configuration. However, the resemblance is not obvious; the back vowels /u ɔ o ɑ a / roughly form a quarter circle shape but /i y/ lie in the periphery, while /œ ø e ɛ / were scattered in the middle of the plot. This configuration and the following remark helps to confirm the suspicion that the stimuli judgments may have been nonlinguistic.

> Since most of the subjects did not even realise that the stimuli were taken from speech sounds, we may assume that they did not use linguistic information in their judgements. In their opinion, they were presented with complex synthetic signals, and they based their decisions on physical cues present in the signals. (p466)

They went on to reject any possible influence from the familiarity of Dutch vowels, by testing the same materials on foreign subjects. Their perceptual and physical structures were also strongly related (correlation values for the first two dimensions range between 0.975 to 0.826). As the number of subjects who had identified the stimuli was the same as the number who took part in the triadic comparison test, one could guess that the

authors used the same subjects for the two different perceptual tests. This point, however, cannot be clarified. If their later remark was correct, then their claim, that the perception of *vowels* can be fully explained by the auditory dimensions emerged from the PCA analysis of spectra, can be undermined, on the grounds that the stimuli used were not heard as vowels, but as meaningless noises.

Also the method of data collection may have helped to preclude linguistic judgements because the ABX similarity judgements could be made "without further indicating the degree of similarity. Moreover, the subjects are not obliged to make their judgements in relation to specific categories."(p460)

### 2.1.2 Klein, Plomp & Pols (1970)

In this study, the above criticisms about speech/noise material and data collection method were removed. 12 Dutch vowels produced by 50 male speakers in an /hVt/ context were used for an identification task. The stimuli were 100 ms of vowel sections extracted from each of the 600 productions. The identification rate of each vowel was accumulated in a confusion matrix (stimuli × response) to be submitted to the Kruskal MDS procedure.

Nevertheless, this identification task, as opposed to the earlier triadic comparison task, is not without its problems. The major problem is that it produces too many zeros in the confusion matrix. This means that many of the vowels were identified as themselves, and not confused with any other vowels. The number of zeros was reduced by enforcing symmetry in the confusion matrix. For example, /ɑ/ was confused with /a/ 38 times while /a/ was confused with /ɑ/ 187 times, producing a total similarity score of 225 (see Table IV in Klein *et al.*). Therefore, the identification method can only be applied to stimuli with some degree of artificiality, in order to bring about confusions. Otherwise there will be an excessive number of equal distances, resulting in an inaccurate estimation of perceptual distances in the map. We return to this point when we are designing the stimuli for the present experiment (See §IV.2).

Klein *et al.* also analysed the same 100 ms vowel sections extracted from each of the 600 (12 vowels × 50 speakers) items by 1/3-octave band pass filtering, 18 filters covering from 100 to 10,000 Hz. The principal component analysis results show 12 clusters representing 50 points for each vowel. The first four components accounted for

about 98 % of the variance, in the case of the 12 average vowel points. The correlation coefficients of perceptual identification to the average vowel values from PCA are very high (0.997, 0.995, 0.974, 0.794).

The resulting three-dimensional perceptual and physical configurations are presented in Figure II-3.

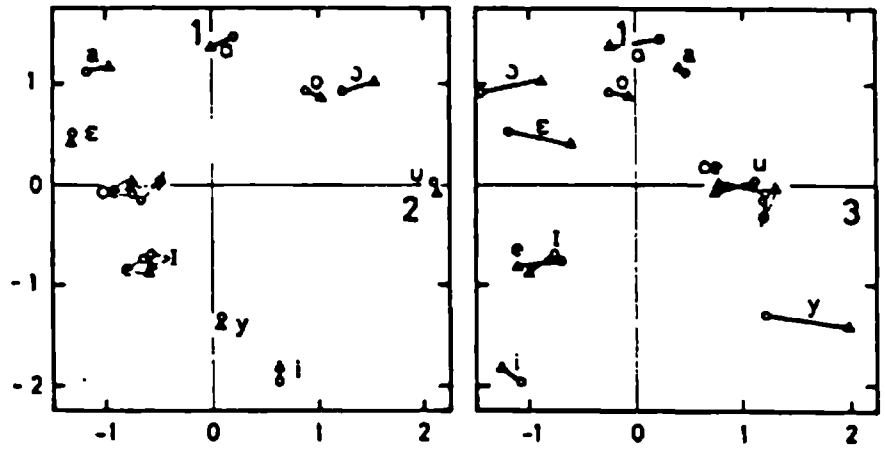



**Figure II-3.** Three most correlating dimensions of perceptual (O) and physical (▲) dimensions after rotating to optimal congruence; after Klein *et al.* (1970).

It is clear from the above figures that the perceptual arrangement of vowels found by Klein *et al.* is more in accordance with the traditional vowel quadrilateral, than was the case with the study by Pols *et al.* (Figure II-2). This is a strong indication that the stimuli were perceived not as vowels, but as noises in Pols *et al.* experiment. On the other hand, the above configuration is not as well correlated with the traditional vowel charts as it was found to be in other studies (See Rakerd & Verbrugge, 1985), and this may be attributed to the inaccuracies involved in reducing the number of zero entries in the confusion matrix, as has just been discussed.

In summary, the most important contribution of these two works, unlike earlier spatial studies on vowels, is that they showed how the auditory space could be directly related to perceptual space. The correlation between the two spaces was high. This was regardless of the modes of perception (Repp, 1982), whether the listeners were listening

to rather unnatural quality of synthetic vowel stimulus or a cutout section of a natural vowel. This suggests that, for the fricative study, we need to be careful about the nature of stimulus types, if we are to make any claims on mechanisms involved in speech perception; we can predict that perception of nonspeech sounds will give high correlation between perceptual and auditory spaces, since there will certainly not be any phonetic influence in the perception of nonspeech sounds.

The following section reviews how any phonetic influence on vowel perception might be studied in terms of spatial configurations.

## 2.2 A study of 'limited' vowel space

Kewley-Port & Atal (1989) applied the MDS technique to investigate an optimal auditory distance metric, to explain the perceptual organisation of vowels in a 'limited' F2/F1 space. The motivation for this work was that there had been numerous studies which tried to model perceptual judgments of speech sounds[1], but that these studies had not applied the MDS technique to transform subjects' distance ratings to spatial representations. Thus, it was hoped that transformations of distances in appropriate MDS spaces might help to evaluate the models of auditory processing.

Kewley-Port & Atal carried out three experiments, and in all cases their stimuli were confined to a subpart of a vowel system rather than using an entire language inventory[2]. In experiment 1, three stimulus sets were synthesised around /i-ɪ/, /e-æ/ and /u-ʊ/ by a LPC digital formant synthesizer. To illustrate, the /i-ɪ/ pair stimuli were such that, on an F2/F1 plane, four vowels were distributed around /ɪ/ and /i/ and one vowel in the middle point between the two vowels (See Figure II-4). The purpose of such stimulus design was to establish the relationship between the perceptual and auditory spaces of the localised region of the vowel space, and also to verify the effectiveness of the MDS technique in reflecting such small perceptual distances. In experiment 2, bilabial bursts and transitions were added to those stimuli in experiment 1, in order to compare the

[1] §3 will be entirely devoted to developments in this area.

[2] We will see in §2.3 that this kind of stimulus design is in accordance with other studies which investigate auditory distance modelling of vowel perception.

perceptual responses of vowels in a consonantal context with the isolated vowel cases. In experiment 3, perceptual distances of eight stimulus vowels were investigated for changes caused by introduction of three distracter vowels. 14 prototypes were generated from the vowel areas /æ-a-ʌ/. The stimuli were divided into two groups of 11 stimuli each, with eight fixed ones and three variable (distracter) ones (See Figure II-5). Comparing the perceptual distance of the eight identical vowels in each stimuli group, it was possible to detect whether those three distracter stimuli in each group had any effect on the perceptual judgements of the vowels in each group.

Listeners' perception of the synthesized vowels was measured by magnitude estimation (ME); point 0 indicated that the pairs were the same, and point 9 meant they are maximally different. This technique is more complex than the triadic comparison test mentioned before (in §2.1) and longer practice sessions were necessary. Subjects had to be familiarised with the stimuli set and given a short practice session at the beginning of each test. Also, the perceptual responses had to be first analysed for consistent use of the ME scale.[3]

K-P & A applied various MDS techniques and found that a 2-way MDS technique produced the most reliable results. The resulting perceptual map was most similar to the Bark scale transform of F1/F2 of the stimuli. Bark scale models the auditory periphery as a bank of critical band-pass filters. This is shown in Figure II-4.

[3]These were the main reasons for rejecting this method of perceptual similarity judgement for the present study (See §3.2).

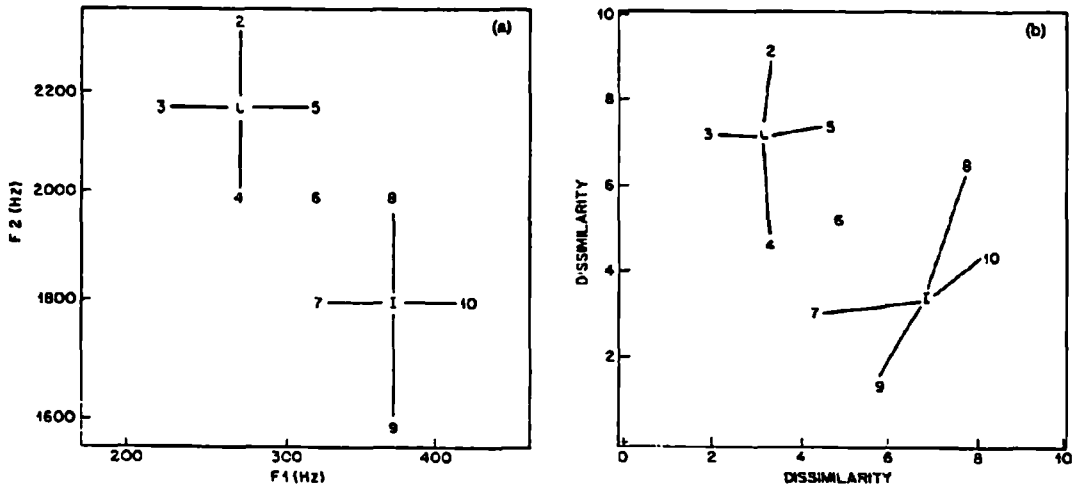



**Figure II-4.** (a) /i-ɪ/ vowels in experiment 1 are plotted on Bark (F2/F1) plane, although labelled in Hz. (b) Two-dimensional MDS solution of the same vowels; after Kewley-Port & Atal (1989).

In order to quantify the relationship between perceptual and auditory spaces, they correlated the Euclidean auditory distances (sum-of-squares of differences of F1 and F2 in Bark between vowels) with inter-pair distances on MDS space (r= 0.94), although canonical correlation analysis (mentioned in §2.1) would have been more appropriate for quantifying the multidimensional correlations between two independent spaces. From this very high correlation, they claimed that:

> ... small differences between synthetic vowels that varied only as a function of two parameters were perceived as varying along only two perceptual dimensions. This was by no means the only possible outcome. Since these vowel stimuli were complex, varying F1 and F2 also caused variation in the overall amplitude, amplitudes of the formants, spectral tilt, and other measurable properties of the stimuli. These properties varied as a direct function of F1 and F2 as defined by the resonant properties of the vocal tract, as modeled by linear prediction analysis. The fact that human listeners did not attend to this variation may be saying something fairly profound about speech perception. That is, the present results seem to argue for the existence of auditory processing mechanisms that incorporate considerable knowledge of the vocal tract. (p1738)

They suggested that this hypothesis can be tested by a comparable experiment with nonspeech material. The perceptual differences between speech and nonspeech materials have been a recurrent problem, as we have seen in the Pols *et al.* study, and we will make

an attempt in §IV.2, to clarify this point. Perhaps more importantly, the problem with investigations involving a subset of a vowel system, is that their perceptual organisation cannot be compared to phonetic organisation, which may point to a particular perceptual mode.

The MDS configurations from experiment 2 were almost identical to the results of experiment 1, and their distance correlations were also very high (r=0.97). This result seems to indicate that the relative perceptual distances were not affected by the consonantal context. The authors attributed this failure to show any contextual effect to the insufficient length of bilabial transition time, and they believe that longer phonetic contexts would have decreased the vowel discriminability.

The results from experiment 3 showed that distance correlations between the two different stimulus groups were very high (r=0.98). This means that, once again, there is no contextual effect in the discrimination of these vowels, but K-P&A claimed that closer examination of the MDS configuration suggests a 'possible' influence of phonetic identity of the stimuli by the presence of the distracter stimuli in the set. This phonetic influence was evidenced by distinct clustering of the stimuli along the perimeter of the MDS configuration which was not observed for the interior part (Figure II-5). According to informal listening tests by trained phoneticians, stimuli on the perimeter tended to be more identified with their prototype vowels than the interior ones.

The implication is that MDS configurations give additional information about the perceptual organisation of sounds, and that this information may be used for identifying more suitable auditory distance metrics, other than Euclidean metric, to model the speech perception. On the face of this result, it is strange that they have not considered applying other distance metrics, which were shown to be phonetically sensitive[4] (See §3.1).

[4]These metrics (reviewed in §3.1) emphasize those spectral parts which are phonetically salient, and deemphasize other spectral parts, which are not important in identifying speech segments (Klatt, 1982a; Assmann & Summerfield, 1989).

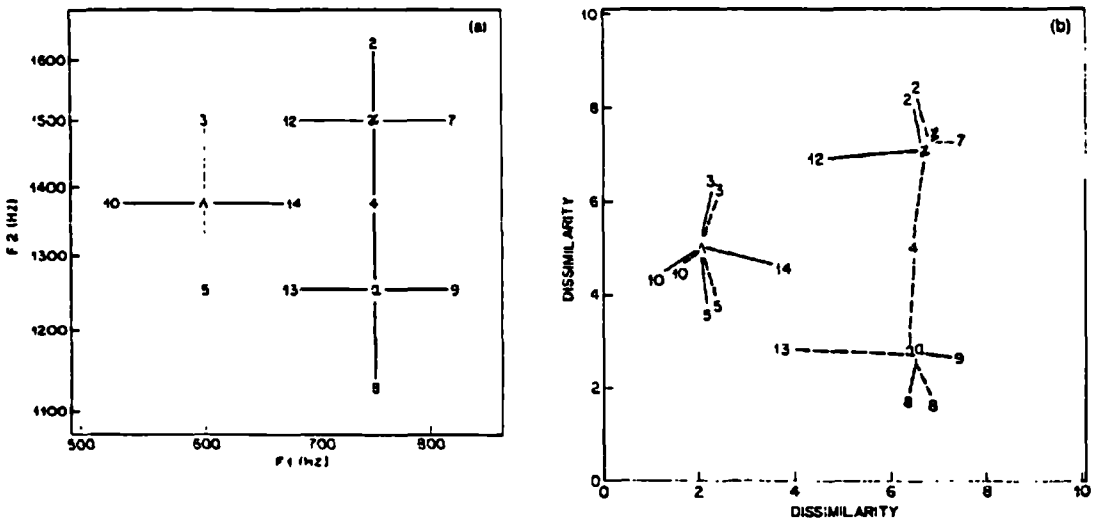



**Figure II-5.** (a) The stimulus set /æ-ɑ-ʌ/ in experiment 3 is plotted on Bark(F2/F1) space, although labelled in Hz. (b) The same vowels are plotted in the two-dimensional perceptual space obtained from the MDS analysis, one stimulus group with solid lines and the other with dashed lines; after Kewley-Port & Atal (1989).

In summary, this study demonstrated the appropriateness of the MDS technique in investigating perceptual relationships between sounds with small physical differences and in various contextual conditions. Although somewhat inconclusive, the study also showed that the technique is also appropriate for showing phonetic influence on the perception of these sounds. On the other hand, we have seen a number of shortcomings in the study, as follows:

(a) they did not use whole vowel space; approach was not holistic;

(b) perceptual and auditory spaces were not directly correlated (only distance correlation analysis was carried out);

(c) auditory distance modelling was limited to Bark scale, and formant frequency specifications;

(d) influence of the phonetic identity of segments on perception could not be properly quantified.

All these points should be taken into consideration in the design of the present investigation.

## 2.3 Acoustic interpretation of Miller & Nicely's data

As mentioned in Klein *et al.* (1970), the use of identification experiments for perceptual research requires some kind of acoustic distortion to induce a sufficient amount of listener confusion. This was the major force in the experimental design of the well-known work of Miller and Nicely (1955). The idea was that the errors induced by acoustic distortion would not be randomly distributed over all the incorrect responses. Instead, consonants that are similar to the intended consonant, in terms of both articulation and acoustics, constitute the bulk of likely errors. It is found that some consonants are more readily confused, that is, they are more similar to one another, than to other consonants. This is tantamount to saying that there are auditory, or articulatory qualities which consonants share to greater or less extent. There have been many attempts to find these common features of consonants based on the confusion data of M&N (Wilson, 1963; Shepard, 1972) and others (Singh & Black, 1966; Singh, Woods & Becker, 1972). However, these failed to recommend a feature system that could account for the perceptual data.

This has led to a reanalysis of M&N's data in terms of possible acoustic features in Soli and Arabie (1979). Further motivation for S&A's reanalysis was that the INDSCAL analysis method had subsequently become available, so that the perceptual influence of 17 different noise conditions in M&N could be fully investigated. Using INDSCAL analysis, the interaction between stimuli × response × noise conditions could be investigated. Thus, acoustic properties of the data are indirectly studied via the listening conditions.

M&N's confusion data were based on English consonants /p t k f θ s ʃ b d g v ð z ʒ m n/, followed by /a/. There were three kinds of listening conditions — noise masking, low-pass filtering and high-pass filtering. At any one time, only one of these conditions was manipulated, while the others were kept constant. This is summarised below:

For the noise masking conditions bandwidth was kept at 200-6500Hz, but the speech-to-noise ratios were varied; 12, 6, 0, -6, -12, and -18 dB.

For the low-pass filtering condition, the speech-to-noise ratio was kept at 12 dB but bandwidths were varied as; 200-5000, 200-2500, 200-1200, 200-600, 200-400, and 200-300 Hz.

Also for the high-pass filtering condition, the speech-to-noise ratio was kept at 12 dB constant, but the bandwidths were varied from 1000-5000, 2000-5000, 2500-5000, 3000-5000, and 4500-5000 Hz.

S&A's INDSCAL analysis of this data showed that four dimensions accounted for 69% of the variation, and the dimensions were best accounted for by the acoustic properties of the stimuli. Each dimension was subdivided into three sections and given appropriate acoustic interpretations. Figure II-6 presents their results.

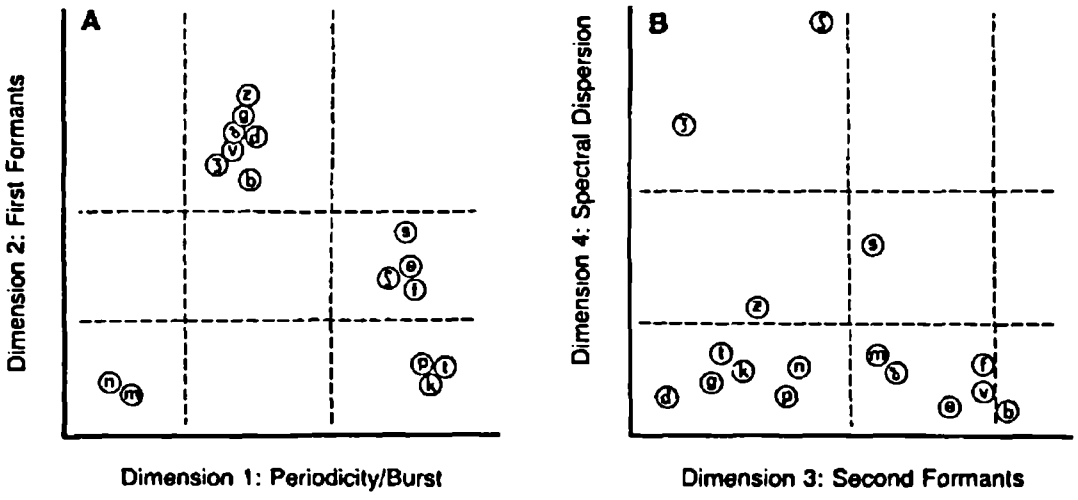



**Figure II-6.** The four-dimensional INDSCAL solution of 16 consonants heard in 17 acoustic disturbance conditions. The broken lines are intended to separate the planes according to their distinguishable acoustic characteristics of syllables; after Soli & Arabie (1979).

Dimension 1 was interpreted according to the acoustic cues 'periodicity/burst'; this was because the first group /n m/ can be characterised as nasal resonance followed by oral release. The second group /b d g v ð z ʒ/ can be characterised by the burst or noise followed by glottal resonance. The members of the third group /p t k f θ s ʃ/ have no

periodicity but they do have burst or noise. However, this dimension can be more conveniently labelled as 'voice'; the first and second groups are voiced consonants and the third group of consonants is unvoiced.

Dimension 2 could not be analysed in linguistic terms; of the three subgroups, the first group /m n p t k/ has a flat first formant; the second group /f s θ ʃ/ has slightly rising first formants; and the third group /b d g v ʒ z ð/ has rising first formants. However, this description would have changed within a different vowel context.

The distribution of dimension 3 showed much less distinctive grouping, but /d ʒ g/ were picked out as having falling second formants, /b/ as having rising second formant, while the remaining consonants had flat second formants. Again this interpretation would change if the following vowels were different; also, this seems to contradict the established importance of F2 as a cue for place of articulation.

The dimension 4 was again segregated into three major groups /ʃ ʒ/; /s z/; and the remainder. This was labelled the 'spectral dispersion' dimension — which is somewhat inappropriate, since the fricatives /f θ v ð/ have spectra with more energy dispersion. Rather, the consonants /s z ʃ ʒ/ are characterised by intense acoustic energy at high frequencies and traditionally grouped together as 'sibilants' or 'stridents'.

On this basis, therefore, the acoustic interpretations seem less than adequate in accounting for the data. Furthermore, no direct matching with the original acoustic data was possible, as M&N had collected the data from spontaneous live productions. Instead, their claim had to be supported by examining the importance or salience of the dimensions with the 'experimental conditions' weights[5] from the INDSCAL solution. The idea is that, if the perception was primarily based on the acoustic cues, then those acoustic conditions, which would degrade the particular part of the spectrum containing the salient cue, would have a noticeable effect on the configurations of the perceptual dimensions.

The dimension weights of the high-pass filtering condition decreased for the first and second formant transition cues and periodicity/burst, while it increased for the spectral dispersion cue. This was accounted for by claiming that "Acoustic information for this dimension is dispersed throughout the high-frequency region and thus is the only

[5]Details are given in §III.2.4.

perceptual information available under high-pass conditions with severe degradation." That is, the spectral dispersion cue was most salient in the bandwidth of 4500-5000; this implies that this spectral region is rich in the salient perceptual cues necessary for the identification of sibilants (on dimension 4). But, in fact, delimiting the analysis frequency to 5000 Hz is not appropriate for the fricative analysis, as will be discussed further in §4. Their explanation is thus not adequately supported.

The dimension weights of noise and low-pass filtering conditions increased for the periodicity/burst and first formant dimensions, but decreased for second formant and spectral dispersion dimensions. This is less problematic to explain, since the second formants occur beyond the range of the some of the low-pass conditions, for example, the bandwidth range of 200-300 Hz.

They also observed that the dimension weight for the first formant dimension of low-pass filtering was maximum between the frequencies 200 and 700 Hz, which corresponds to the frequency ranges in which the first formant transitions of female speakers usually occur. For the second formant dimension the weight decreased most rapidly for frequencies between 1000 and 2500 Hz, which again corresponds to the second formant transitions ranges.

In an attempt to give an even more detailed reanalysis of M&N data, Soli and Arabie, with Carroll (Soli, Arabie & Carroll, 1986) conducted an Individual Differences Clustering analysis. This method recognises finer perceptual divisions between consonant classes as shown in Figure II-7 (the actual configurations of consonants on the four dimensions are exactly the same as in Figure II-6), and each cluster or division may be compared with the acoustic conditions. This is done in terms of the contribution of individual cluster weight for estimated similarity of consonants in that particular cluster. For example, /d g/ occurs in clusters I, V, XII, and XIII. The estimated similarities of /d g/ is modelled as a sum of all the cluster weights in which /d g/ pair occur and a specified constant. The estimated similarities for /d g/ in the low-pass conditions (400-5000 Hz) were almost entirely accounted for by the weights in cluster I. This implies that important spectral information concerning the place cue is contained in this region. For the low pass condition of 200 to 400 Hz, the weight from the other cluster V contributed significantly to the estimated similarity of /d g/. This implies that in frequency bands 200-

300 Hz, the place cue does not exist. In the same way, the estimated similarities for /p k/ were calculated in low- and high-pass conditions. The results indicated that place of articulation cues were not confined to the spectral region of the second formant, but spread out across the frequency region between 400 and 5000 Hz.

Although the same type of analysis has shown that high frequency spectral energy is an important cue in fricative perception, the observation of any direct relationship between perceptual and acoustic properties was beyond the scope of such an analysis, due to the inherent lack of acoustic data in the original M&N experiment.

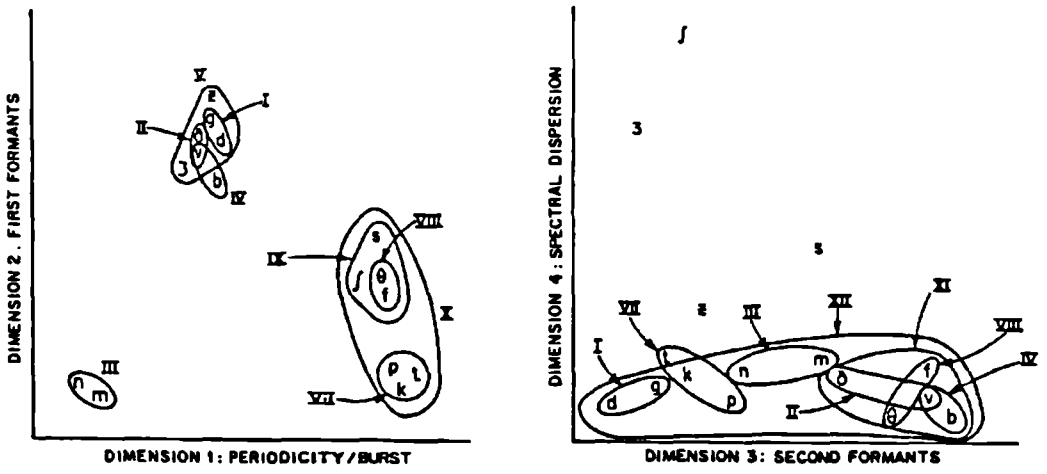



**Figure II-7.** INCLUS clusters comprised of voiced consonants (I-VI), voiceless consonants (VII-X), and mixed-voiced clusters XI and XII displayed in the plane of (a) INDSCAL dimensions 1 and 2, (b) INDSCAL dimensions 3 and 4, in Figure II-6; after Soli, Arabie & Carroll (1986).

In summary, some useful insights into the relationship between perceptual dimensions and acoustic conditions have been formulated in these papers. This type of study could provide a detailed representation of the perceptual structure of consonants, and how it is related to the acoustic degradation conditions; however, these studies have generated rather speculative relationships. Much more work was required in order to reach a similar degree of correspondence between the perceptual and acoustic domains of consonants as has already been achieved in vowels. Nevertheless, this type of study has been scarce. One such study, involving the three-dimensional acoustic space of voiced stop consonants, is reviewed below.

## 2.4 Spatial representation of stop consonants

Sussman (1991) represented English initial voiced stop consonants /b d g/, produced in ten different vowel contexts in a three-dimensional acoustic space. The motivation for this work stems from difficulties in accepting *absolute* invariance for stop place, independently of vowel context (Stevens &Blumstein, 1979). The difficulty was not so much in the specification of the absolute acoustic patterns themselves, but in perceptual verifications; vowel contexts still conditioned perceptual results (p19). Instead, Sussman suggested a form of *relational* invariance, following Lindblom (1990). According to this view, each stop place is represented as a distinct (nonoverlapping) region on an acoustic space.

The acoustic space was based on multiple tokens of /bVt/, /dVt/, and /gVt/, with ten different medial vowels, produced by ten male and ten female speakers. On the basis of the analysis by Lindblom (1990), he identified the three most important acoustic parameters: second-formant onset frequency (F2 onset), third-formant onset frequency (F3 onset), and second-formant frequency of the vowel nucleus (F2 vowel). These correspond to the three coordinates of the acoustic space. The three-dimensional acoustic space of the ten male speakers is presented in Figure II-8.

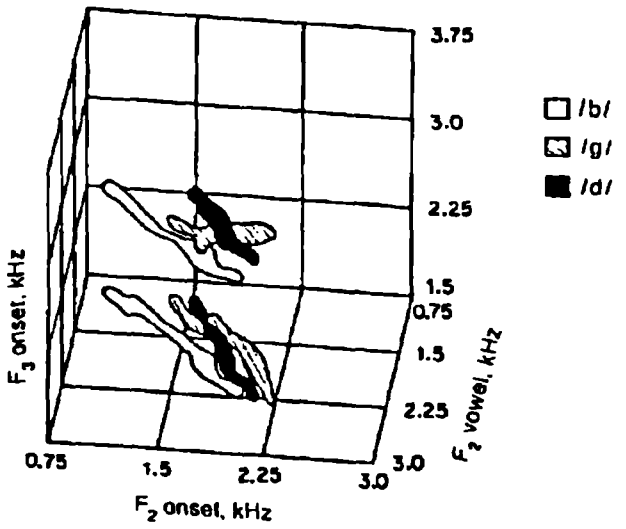



**Figure II-8.** A three-dimensional acoustic space of voiced stop consonants /b d g/. Male grand mean (n=10); after Sussman (1991).

The regions corresponding to each stop are enclosed by "cloud" outlines, hovering above in three-dimensional space, and their "shadows" or "depth dimension" are shown on F2 vowel and F2 onset axes. None of the regions are overlapping. Additional discriminant analyses, based on the three acoustic parameters, showed an average correct classification rate of 84.5%.

From these results, Sussman (1991) formulated a kind of spatial isomorphism between physical properties and linguistic representations:

> Topographical mapping of sensory systems reflects an explicit isomorphism between the spatial geometry of the receptor surface, i.e., retina, basilar membrane, skin/tissue, and the neural representation (retinotopic, tonotopic, somatotopic, respectively) of these surfaces. This isomorphism between receptor surface and brain mapping allows meaningful decoding of the sensory stimulation pattern. ... Neural representations of speech should be expected to conform to the physical patterns of the signal capable of sufficiently contrasting phonemic categories. Selected information-bearing elements contained in the sound waveform can be regarded as neurally mapped acoustic variables. In such a view the phonological abstraction of the phoneme achieves, in principle, a hard-wired instantiation, i.e. a 'spectrotopic' representation. (p30)

This kind of assumption is not altogether unfeasible, but there are some shortcomings in drawing such a conclusion. First of all, Sussman represented three phonetic categories in a three dimensional acoustic space. An obvious question arises: 'Do we need 20 dimensions to represent 20 consonants?' Surely the representation of 20 objects on 20 dimensions will always result in the perfect separation of each object. What we clearly need is a reduction in the number of dimensions required to represent the data. A possible solution to this problem may be the principal component analysis (PCA), mentioned earlier in Pols *et al.* (in §2.1).

Another problem with Sussman's acoustic space lies with the actual specification of the acoustic coordinates. The formant onset values are the major parameters in this representation. But it is established that the formant onsets may not be so easy to identify in voiceless stops, and therefore, we may need an entirely new set of parameters to represent the place categories in the voiceless stops. Even for the voiced stops, Sussman used three different types of analysis to determine the formant frequencies. Again this kind of *a priori* assumption can be avoided if we are able to do a whole spectral processing,

as in PCA.

In addition, the perceptual tests to verify the perceptual significance of these acoustic dimensions were not carried out.

## 3 Modelling of perceptual distance judgements

### 3.1 Distance metrics

#### 3.1.0 Introduction

The studies reviewed in the previous section used Euclidean distance metric to account for auditory spectral differences. In Pols *et al.* (1969) and Klein *et al.* (1970), the distances between vowels on the auditory (PCA) space were the sum-of-squares of the differences in critical-band filter bank outputs.

Yet it has been suggested repeatedly in the literature, that we need to devise a metric which would be more selective in its sensitivity and give more weight to those spectral properties that are perceptually salient. One of the best known works of this kind is by Klatt (1982a). The metric suggested by Klatt was further improved by Assmann & Summerfield (1989). In both of these works, the stimulus sets have always been vowels. These two works, as well as two consonant distance studies, are reviewed here.

#### 3.1.1 Weighted slope metrics

This section is concerned with a review of the well-known phonetic distance metric analysis by Klatt (1982a). There are two major parts to this investigation. Firstly, Klatt conducted three separate perceptual experiments to identify the spectral properties which were directly related to the phonetic identity of a segment. However, it is the third experiment which is of particular interest to us here, since the experimental materials were extended to include fricative-like sounds. This indicates a strong possibility that the metrics suggested in the second part may work equally well for the fricative analyses. This latter part is a proposal for a suitable distance metric to account for those spectral aspects which were identified as phonetically significant.

The first experiment was based on the same set of stimuli as was employed in

Carlson, Granstrom & Klatt (1979), in which this kind of phonetic distance modelling has its origins. The materials for the first experiment were 66 variations of /æ/ produced by harmonic manipulation. Formant frequencies which were thought to be phonetically important, and 'non-phonetically' important parameters such as spectral tilt changes, relative formant amplitude differences, high-, low-pass and notch filtering, were all tested out. The subjects' task was 0-10 point scaling of pairwise similarity judgements. There were 300 trials for 8 subjects (how this was arranged is not clear). Subjects were asked either to i) take into account any differences between the vowels (*psychoacoustic differences*) or ii) rate only changes that tended to influence vowel identity (*phonetic differences*). The normalised subjective perceptual distance was related to changes in each acoustic property.

It was clear from the normalised distance ratings that psychoacoustic and phonetic judgements were different. Only the formant frequency changes induced large phonetic distance changes, though the psychoacoustic distances changed according to various other parameters. This can be clearly shown by the first three parametric conditions, as shown below:

| | (i) Average Psychoacoustic distance | (ii) Average Phonetic distance |
|---|---|---|
| Random phase | 20.0 | 2.9 |
| High-pass filtered at 300 Hz | 15.8 | 1.6 |
| F1, F2, F3, F4 & F5, +8, -8% | 10.7 | 8.1 |

**Table II-1.** Comparison of psychophysical and phonetic distance judgments for the voiced /æ/ stimuli, for the first three synthetic conditions; after Klatt (1982a).

This result was enhanced by experiments 2 and 3. Experiment 2 used /a/ vowels instead of /æ/, in order to test whether the formants, when they are very close together, would still show a high correlation with the phonetic judgements (in particular, the salience of formants in the vowel perception). In experiment 3, noise excitation /æ/ vowel was investigated. The overall result shows a tendency similar to that found in experiment 1. In a further vowel-like or fricative-like distinction test, the stimuli which had higher frequency formants than the reference tended to be judged as fricatives. Also, the spectral

tilt information had greater significance for voiceless sounds, "possibly because of the nearby phonetic boundary between vowel-like and fricative-like". This seems to imply that the fricatives can be analysed as noise excited vowels with high frequency formants (for further discussion, see §4), and the distance metrics suitable for vowel analyses may also be appropriate to the fricatives.

The second part of this study was devoted to the development of a distance metric which computed phonetic distance differences, without any formant extraction procedure. Lindblom (1978), Carlson & Granstrom (1979) and Bladon & Lindblom (1981) had shown that the psychophysical distances can be suitably accounted for by Euclidean distance metric applied to critical-band filterbank analyses (correlations between perceptual distances and Euclidean metric based auditory distances were ~0.85). However, Klatt showed that the Euclidean metric works less well for stimuli in which there were large spectral differences but very few phonetically significant changes. To tackle this problem, Klatt developed the **weighted slope metric (WSM)**, which takes the first differential of the spectra, so that it is sensitive to formant frequency values, but not to relative amplitudes nor to the changes in low energy regions. Spectra were first analysed by 36-channel bandpass analyses. The actual formulae and further details of the metric is discussed in next chapter (§III.3.2.3).

The best prediction of phonetic judgement of /a/ from the new spectral slope metric was 0.93, which was significantly better than previous attempts. As already mentioned in §2.2, therefore, it would be worthwhile applying this metric in the study of auditory spaces.

This metric was also tested out for speech recognition. Nocerino *et al.* (1985) have applied several metrics in speaker-specific, isolated word speech recognition tasks, and reported best results for this spectral slope metric.

### 3.1.2 Weighted negative second differential (N2D) metric

In another vowel study, Assmann and Summerfield (1989) furthered the distance metric modelling of vowel perception. However their experimental design was markedly different; they investigated the perceptual robustness of formant frequency information in the presence of a competing 'voice', using judgements of so-called double vowels.

The materials were designed in three different groups:

Group 1. Five formant cascade synthesis vowels based on /a i ɔ ɜ u/

Group 2. 'Six-harmonic' vowels, in which pairs of harmonics of equal amplitude were arithmetically centred on the centre frequency of the lowest three formants.

Group 3. 'Six-harmonic' vowels 9 dB above a competing spectrum of uniform harmonics.

Vowels in each group were paired to make up the 'double vowels'. Identification judgements of each constituent of double vowels were carried out. The 'cascade' and 'six-harmonic' vowels were perceived in a similar way. This was interpreted as confirmatory evidence for the importance of formants in vowel perception.

The auditory modelling of the subjects' responses was rather complicated, but the main assumption was that vowel perception is a process of pattern matching, so that the 'double vowels' were treated as target patterns and single vowels as reference patterns. First, auditory frequency analysis was based on 512-point FFT on vowel centre slice, and it was converted to 300-point excitation pattern. Notice that, in comparison to Klatt's study, this analysis is based on energy outputs from 300 channels. Second, they converted auditory excitation patterns to 'target patterns' by four different representations — energy levels, slopes, N2D, and peak picking. The N2D metric only incorporates the negative part of the second differential of the spectral level, assigning a zero value for the positive part. In effect, this metric emphasises spectral formants, while ignoring other properties of spectra completely. Thus, N2D was a step beyond Klatt's metric in emphasising formants (See §III.3.2.3). The reference patterns were generated via the same processes from the single vowel waveforms instead of the double vowels. Lastly, distances between the target and reference patterns were compared using four different distance metrics.

They reported that the N2D metric performed best out of the four metrics. Although this metric has not been tested on other vowel stimuli, this provides an incentive for applying N2D metric, along with Klatt's, for the present fricative study. A difficulty will be to account for the dynamic spectral changes in fricatives, rather than taking the

centre slice from steady-state vowel spectra. We will describe in detail how this problem can be solved in detail in §III.3.2.2. Before suggesting our own solution, we review other people's attempts at solving this problem in the next section, with regard to stop consonant perception. We have not encountered acoustic distance modelling for fricatives so far in phonetic science.

## 3.2 Some attempts at auditory distance modelling for consonants

### 3.2.0 Introduction

In this section, two studies which attempt to model the perception of stop consonants in terms of their auditory spectral distances are reviewed. As mentioned earlier, the challenge in auditory distance calculations for stops is to quantify the rapid spectral changes. Sidwell & Summerfield (1986) used perfectly time aligned spectra, while Krull (1990) used *ad hoc* combinations of various acoustic properties to account for this dynamic change in the stop spectra. These shortcomings point to an alternative technique for time-alignment of spectra and a more generalised method of auditory distance analysis.

### 3.2.1 Uniform alignment of spectra

Sidwell & Summerfield (1986) extended the scope of auditory distance modelling studies from steady state vowels to time-varying auditory spectra. They used 'voiced' stop consonant syllables, /bəb dəd gəg/, which were 'whispered', so that the stimuli had clear symmetrical initial and final formant transitions, but without the release burst. More specifically, they synthesized a vowel with formants at 500, 1500, 2500 and 3500 Hz and appended the formant transitions appropriate to each of the places of articulation of the stops, on both sides of the vowels. Schematic spectrograms of the stimuli are shown in Figure II-9.

Then the CVC syllables were divided into CV and VC syllables with an extended vowel part to make a total length of 200 ms. The listeners were asked to discriminate between the pairs /b-d, d-g, b-g/ in both syllable-initial and syllable-final positions.

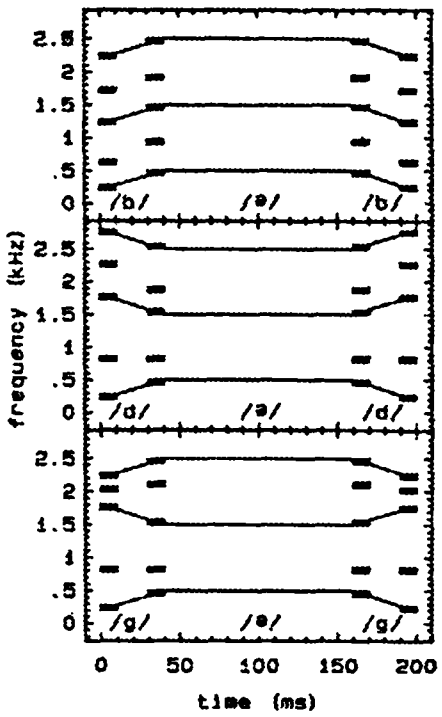



**Figure II-9.** Schematic spectrograms of the symmetrical CVC syllables. Each syllable also contained a stationary fourth formant at 3.5 kHz. The solid rectangles at the onsets and offsets of transitions represent the locations of the 10 ms sinusoidal signals used to measure the masking patterns of the stimuli; after Sidwell & Summerfield (1986).

Because the spectra are perfectly aligned, the auditory distances could be measured by comparing the "masking patterns" (taken to reflect the auditory spectral shapes) at particular points in time. They selected four sampling points; formant transition onset and offset in syllable-initial and syllable-final positions (as indicated by the solid rectangles in Figure II-9). They then carried out masking pattern analysis[6] at seven predetermined frequency points, corresponding to peaks and valleys of the stimulus spectra.

Auditory distances between the pairs /b-d, d-g, b-g/ were calculated by taking a 'peak' or 'total' difference in masking patterns along the seven frequency points. Also, peak and total differences were calculated over the 'formant transition' interval.

The best correlation found between the discriminability scores and the masking patterns, were of the peak differences at onset/offset, and the total transition differences,

[6]For a detailed description of this analysis, see p288 in Sidwell & Summerfield (1986).

at r=0.89 and r=0.94 respectively.

In this study, the auditory distance analyses were made possible only because the compared stimulus pairs were perfectly time-aligned. In real speech, however, we must account for differences in speaking rates and voice onset time in different stop consonants. For example, the voice onset time of /g/ is usually longer than that of /b/, thus the onset and offset of transitions are rarely aligned. In view of these problems, we must first find a means of aligning the stimulus spectra in a way which matches the corresponding points of the equivalent part of the spectra, before masking patterns are compared. Another problem with the distance calculation in this study was the excessive subject time required to measure the masking patterns; measurements of the masking filter took about 15 hours of subject time per masker. Thus, the effects of each masking pattern were calculated from the judgments of a single subject (total number of subjects was seven). This may have contributed to the inaccuracies in calculating auditory distances.

Overall, these methods are clearly inappropriate for any large scale investigation of fricatives.

### 3.2.2 'Burst length'

Another study which attempts to model the perceptual confusion of stop consonants by acoustic/auditory distances was undertaken by Krull (1990). She used natural speech, consisting of four intervocalic voiced Swedish stops in five different vowel contexts. The stimuli were cut into VC, CV, and 26 ms consonant release sections, and used for confusion tests.

The confusion data were predicted from the acoustic distance analyses based on a combination of *static* and *time-varying* auditory distances. The static distances were measured in terms of:

i) sum-of-squares of F2, F3 and F4 at the CV boundary measured in Bark and;
ii) spectral level differences from 14 filter, 1/4-octave band pass filtering, at 10 ms after consonant release.

Each static distance was combined with different time-varying properties:

- For (i), time-varying distances were measured in terms of 'burst length' (voice onset time).
- For (ii), time-varying distances were measured in terms of the spectral level differences between 10 ms and 20 ms after release burst.

The above choices for time-varying distances were based on earlier experimentation, which had found that the voice onset time of stop consonants varies with place of articulation, and that acoustic invariance was identified in the spectrum shape during the 10-20 ms following stop release (Stevens & Blumstein, 1979).

Krull reported that the best prediction for the confusion data was obtained with measurements of formant values at 10 ms after the burst, in combination with burst length data (Spearman rank-order correlation between perceptual confusions and auditory distances was 0.85).

However, this result was driven by auditory analyses which were based on a preconceived notion of the importance of particular combinations of acoustic properties. In order to obtain a reasonable correlation, many different combinations of given properties were tested. For example, *ad hoc* weighting factors were introduced, as illustrated with the following formula:

$$D_{mij} = [(w_1 D_{formij})^p + (w_2 D_{bj})^p]^{1/p}$$

where $D_m$ is the combined distance, $D_{form}$ is the sum-of-squared differences between F2, F3 and F4 in Bark, and $D_b$ is the difference in burst length in ms, in *i* and *j* stimuli. $w_1$ and $w_2$ were weighting factors, set to 1.0 and 0.1 respectively. On the other hand, in the calculation of the spectrum-based distances, the static and time-varying distance measures were simply added together without weighting factors. The crucial point here is that, by following this procedure, weighting factors can be introduced *ad hoc* and manipulated until a reasonable level of correlation is reached. This kind of hypothesis-driven manipulation of acoustic components leads to countless alternative analyses, which are precisely what we wish to avoid.

In summary, we have shown in this section that, Sidwell & Summerfield's auditory distance analysis was limited to the strictly time-aligned synthetic materials,

while Krull's analyses were based on *ad hoc* combinations of acoustic properties. In view of these problems, the present study will endeavour to employ more general methods of auditory distance analyses, by applying a *non-linear time alignment* technique (see §III.3.2.2), before critical-band filter bank outputs are analysed using different distance metrics (see §III.3.2.1), and represented in a geometrical model (in §III.3.2.4).

In the following sections, previous studies of fricative acoustics are reviewed.

## 4 Acoustic cues for fricative perception

### 4.0 Introduction

The objective in this study is to investigate how closely the acoustic/auditory space for consonants can mirror their perceptual space. We have seen that the perceptual space of vowels is simply related to acoustic/auditory space (in §2.1). For consonants, this kind of spatial relationship has not yet been established. The comparison between auditory distance analyses and corresponding perceptual judgements in consonants has been limited to stop consonants (§3.2), and the two studies reviewed pointed out some of the problems involved in auditory modelling of consonant perception.

Pursuing this line of study, fricatives or fricative-like sounds provide a convenient medium in which the knowledge of vowel auditory modelling can be applied to auditory analyses of consonants. Some examples of auditory modelling of fricative perception have already been mentioned: Klatt (1982a) used /a/-like vowels with noise excitation, and Sidwell & Summerfield (1989) had used whispered voiced stops without the stop bursts. The present investigation is concerned with fricative consonants, for which five places of articulation can be identified in standard English. The five voiceless fricatives /f θ s ʃ h/ should be sufficient in number to provide a coherent group of objects on an acoustic space[7].

According to the traditional *source-filter* acoustic model of vowels, fricative excitation is made by a narrow oral constriction; when air passes through the narrow gap, a turbulent noise excitation is caused. A more recent model of fricatives is presented by Shadle (1990). According to her model of fricative production, there are at least two major sources of fricative excitation. The first is an *obstacle source*, according to which

[7]In the main perceptual experiment, we have also added the fricative /x/. See §IV.2 for details.

the sound is generated at a rigid body downstream from the airflow. /s/ and /ʃ/ are examples of the fricative sounds generated by this source at the teeth. The second source is a *wall source*, in which case excitation is generated along a rigid wall that is parallel to the airflow from eddies in the boundary layer of the flow. Examples are the fricatives /ç/ and /x/.

Fricative production is usually modelled in terms of *poles* and *zeros*. The poles are the natural resonance frequencies of the vocal tract, and they do not depend on the location of the excitation source. The zeros are anti-resonances, and, according to Shadle's model, they are mainly caused by the irregular shape of the anterior cavity. Because the constriction is so great, anti-resonances from the posterior cavity have, in fact, negligible effect. (for details, see Shadle, 1990).

Consequently, fricative spectra can be characterised in terms of poles and zeros, as is the case for vowel spectra. These spectral properties can be represented as relative positions of fricatives on their acoustic/auditory space.

Despite this possibility, we have not yet come across any spatial analysis of fricative acoustics, nor fricative perception. As is normally the case with acoustic studies, the fricative studies have mainly centred on the investigation of perceptually salient cues. These detailed cue studies usually lead to futile circular arguments, without offering any insights into the speech processing mechanism in general (refer to §I.2). However, these studies provide important information concerning the acoustic properties of fricatives, and thus, need to be referenced in the stimulus design of present study.

In the following subsections, we give a brief overview of acoustic cue studies concerned with different places of articulation in fricatives, in terms of i) amplitude-, ii) spectral peak-, iii) adjacent vowel environment-, and iv) duration-cues.

### 4.1 Frication amplitude cue

Traditionally, fricatives have been classified in terms of their amplitude property; the stridents or sibilants /s ʃ/ are characterised by greater acoustic energy in the noise spectrum than non-stridents /f θ/. Because of this intense acoustic energy, the perception of stridents was thought to be primarily based on the fricative noise section. The earliest experiment demonstrating this point was conducted by Harris (1958). She used a splicing

technique to augment conflicting vocalic transition sections to the fricative noises, in order to evaluate the perceptual importance of friction noise and transition. Three major conclusions could be drawn from the result: i) that "the friction of /s/ and /ʃ/ provide the necessary and sufficient cues for their identification, and override whatever cues may be proved by the vocalic portion", ii) that the /f θ/ "class" was completely distinguishable from the /s ʃ/ class on the basis of friction, but iii) the vocalic portion was necessary to distinguish between the non-stridents, /f/ and /θ/.

LaRiviere *et al.* (1975) also studied the amplitude properties of frication noise. Their results agreed with those of Harris, in that strident perception was primarily based on the frication noise spectra. However, there was further evidence to suggest that the vocalic transitions and vowel context influenced the perception of /s ʃ/. The perceptual effect of vocalic portions will be discussed in §4.3.

The significance of overall amplitude cue has been contradicted by Behrens & Blumstein (1988). They have found that, when both frication spectrum and formant transition cues were appropriate for a specific fricative, the perception of fricatives was not affected by the overall amplitude of frication. Instead they have suggested that the *relative* amplitude of frication in relation to the vowels in a specific frequency region may be the relevant amplitude property for fricative perception. Stevens (1985) also showed that the relative amplitude in a particular frequency region could influence the listeners' perception of /s/-/ʃ/ and /s/-/θ/ contrasts.

However, the relative amplitude variations, in relation to other cues such as spectral peaks, formant transitions and duration, was shown to be a *secondary* perceptual cue, and only used when the perception from the *primary* cues was obscured (Hedrick & Ohde, 1993). This notion of primary/secondary cues follows from the acoustic invariance theory (Stevens & Blumstein, 1979).

In a study of the Spanish fricatives /f s ʃ x/ by Manrique & Massone (1981), where /θ/ was absent, the binary distinction between stridents versus non-stridents was not treated with any major significance. They concluded that the filtering and synthesis experiments showed that:

> ...the main concentration peak carries the perceptual load for the identification of [s], [ʃ], and [x]. In the case of [f], its relevant characteristic seems to lie in a

diffuse spectrum. (p1152)

This leads to the question of the importance of the spectral formant cue in the perception of fricatives.

### 4.2 Spectral formant cue

An important cue for distinguishing between the fricatives /f θ s ʃ h/ lies in variation of energy distribution in the fricative noise portion. Traditional studies on fricatives show that /s/ and /ʃ/ have spectral peaks typically at 4-8 kHz for /s/ and at 2-4 kHz for /ʃ/, whereas /f/ and /θ/ have an even distribution of energy in this range. In the case of the glottal fricative /h/, the frequency of the energy peak follows the formant values of surrounding vowels.

In a perceptual study of the fricatives, Heinz & Stevens (1961) synthesised fricatives by filtering the white noise through spectra consisting of a single resonance (a spectral pole) and a single anti-resonance (a spectral zero). Perceptual responses were such that the noises with the centre frequency of the resonance below about 3 kHz were perceived as /ʃ/ and the centre frequency between 4-6.5 kHz as /s/. Resonance frequencies around 6.5 or 8 kHz yielded /f/ and /θ/ responses. Low overall intensity level of frication relative to the vowel portion helped the response of /f θ/. Although no concrete experimental evidence was presented, they have inferred from the stop consonant perceptual experiments that the /f-θ/ distinction would be made on the basis of F2 transition cue.

This result was further clarified in a study by Manrique & Massone (cited earlier) on Spanish fricatives. White noise was filtered through various high- and low-pass filters. Corresponding frequency bands responsible for /s/ identification were between 5-8 kHz, and /ʃ/ identification was related to a peak frequency of about 2.5 kHz. For the identification of Spanish /f/, both the low, as well as the high frequency zones (below 1000 and above 7000 Hz respectively) were necessary.

In addition to the spectral formant cue, surrounding vocalic contexts may also play a prominent role in fricative perception. Perceptual importance of vocalic contexts forms the topic of next section.

### 4.3 Vowel quality and formant transition cues

Contrary to the assumption that the vocalic transitions play a prominent role in the distinction of the fricatives /f-θ/, Delattre *et al.* (1963) showed that the vocalic transitions do not carry important perceptual cues for the identification of the fricatives /f θ/. To maximise the effect of transition cue they used voiced fricatives, and for the /v-ð/ pair, broad band noise was used with various second and third formant transitions. The perception of these synthesised stimuli showed some effect of transition, but the transition cue, in various vowel environments, failed to "provide the basis for unequivocal discrimination" (p113). However, their second experiment with straight formants (zero transitions) produced clear target frequencies, according to which the fricatives /f-θ/ were categorically perceived.

LaRiviere *et al.* (1975) showed that adding the vocalic transitions to /f/ and /θ/ did not improve identification scores. Furthermore, they have shown that the vowel segments added to the initial fricatives, without the transitions, improved the perception of the fricatives in the sequences /fa, fu, θi/. But for /θa/ segment, adding neither the vowels nor the transitions improved the perceptual scores. For /fi, su, si, ʃu/, transitionless syllables resulted in higher identification scores and no context effect of the vowels was reported.

Manrique & Massone (1981) speculated that the results in LaRiviere *et al.* followed from the frication and vocalic portions being "independently produced", while in their own experiment both proportions were gated out from the same natural syllable, thus the perceptual effect of transition was observable. They suggested possible "backward coarticulatory influence" in the frication portions, "which prevents the vowel effect to arise". They also observed that when the frication and vocalic portions were compatible the identification score had increased, but when the vocalic part was in conflict with the frication portion, the score had decreased. This indicates that the vocalic effect on the perception of the fricatives may be secondary.

Mann & Repp (1980) and Mann & Soli (1991) showed that the distinction along the [ʃ]-[s] continuum was strongly influenced by the formant transitions and by vowel quality. Mann & Soli (1991) showed that an /u/ vocalic context induced more /s/ identification than /ʃ/. No general consensus, however, has been reached with regard to vowel context. Hedrick & Ohde (1993) report that a following /u/ vowel, compared to

/a/ and /i/, resulted in fewer identifications of /s/ in /s/-/ʃ/ than in the /s/-/θ/ contrast.

Since the evidence for perceptual importance of vowel quality and formant transition cues is inconsistent, the effects of these properties are also investigated in the present study, especially the perceptual effect of transition cues.

### 4.4 Frication duration cue

Fricatives are like vowels to the extent that they are continuants and their sound can be prolonged. The temporal aspect of their spectra on perception has not been investigated, except on the influence of the duration factor. Differences in the duration of fricative noise as cues to place of articulation have been investigated by, among others, Manrique & Massone (1981) and Jongman (1989). It is generally established that the average duration of fricatives in English increases according to the following category order: dentals, labials, alveolars and palatals. (This is in agreement with the findings in Manrique & Massone on Spanish fricatives.)

In pursuit of the acoustic invariance theory (Stevens & Blumstein, 1979), Jongman claimed that fricatives vary greatly in the stimulus duration that is needed for correct identification, and suggested that more than 40 ms of fricative segment is necessary to build up to the amplitude level sufficient for identification. The result showed that only 30 ms was sufficient for the identification of /ʃ, z/, whereas /f s v/ needed about 50 ms for a comparable rate of identification. The fricatives /θ, ð/ were identified reasonably well only when the whole section of frication was given. However, it could not be determined from the study whether the dynamic changes within the duration interval had any perceptual significance, or whether there is a critical time interval (or a particular spectral slice) which contained major perceptual cues.

In summary, we have seen that several kinds of cues interact with each other to specify any single fricative contrast. We shall investigate the effects of these acoustic properties on perceptual, auditory, and phonetic spaces.

## 5 Summary

In this chapter, we have reviewed studies which have directly related the perceptual and auditory spaces of vowels and vowel-like sounds. As expected, the correlations were very

high. On the other hand, we have seen that equivalent studies concerned with consonants have failed to establish such a simple relationship between the physical and perceptual domains.

We have also reviewed auditory distance metric modelling of perceptual judgments for vowels, and commented on the possibility of implementing such metrics for fricative analyses. In addition some of the stop consonant acoustic distance analyses were also reviewed.

Finally, basic articulatory and acoustic cue studies of fricatives were briefly discussed. According to these studies, it is possible that many different kinds of cues are interacting to specify any one perceptual category; the cues are absolute and relative amplitude, surrounding vocalic contexts, spectral formants, and duration. These acoustic properties will be referred to in the design of present perceptual experiments and the subsequent auditory analyses.

# *Chapter III. Preliminary analyses on phonetic, perceptual and auditory spaces of fricative sounds*

## 1 Introduction and objectives

This chapter is concerned with preliminary analyses on phonetic, perceptual, and auditory spaces for a coherent group of consonants in English: the voiceless fricatives. Fricatives are chosen because: (i) in theory, at least, they can be modelled in the same way as vowels, in terms of the spectral poles and zeros, as described in §II.4; (ii) they occur in five different places of articulation in English, and this will give a larger number of combinations of similarity judgements and points on the perceptual space. Moreover, the place of articulation feature of consonants is chosen because, conventionally, it is recognised as the major *phonetic* criterion. Also the *perceptual* reality of the place feature is well established in previous studies (e.g. Singh, Woods & Becker, 1972), and furthermore, the *acoustic* analysis for place cues has been most thoroughly studied. Only voiceless fricatives are used, so that the acoustic parameters involved are relatively homogenous.

The preliminary analyses presented here were necessary because: 1) unlike the case of vowels, the phonetic and perceptual spaces of fricatives are not clearly established; 2) the relationship between perceptual and auditory spaces of fricatives has never been studied; 3) if perceptual organisation is not explainable by auditory modelling, we need to develop appropriate methods for MDS experiments to indicate any phonetic influence on perceptual similarity judgments. Three separate preliminary studies are carried out here, and brief summaries of each are given below.

Experiment 1 (in §2) was set up to establish the perceptual map of natural fricatives in the context of a following /ɑ/ vowel, so that its perceptual configurations could be used as reference measures for the synthetic stimulus signals. The major dimensions to emerge from the MDS analyses were 'place of articulation' and 'sibilance'.

Experiment 2 (in §3) was made up of two separate sets of stimuli: the first set was made up of fricative sections only (as in Klein *et al.* study, §II.2.1.2), cut out from the original stimuli in experiment 1; and the second set was LPC synthesised fricatives

modelled on the same stimulus set (as in Pols *et al.*, §II.2.1.1). Distance modelling of perceptual configurations of these stimulus sets showed that correspondence between phonetic, perceptual, and auditory spaces were not as straightforward as for vowels. A poor correlation between these spaces suggests the need for more detailed analyses of fricative data.

In experiment 3, the perceptual and auditory spaces of two formant white noises were investigated to verify the hypothesis that, for meaningless noises, perceptual organisation should be completely explained by auditory organisations. The result showed that the Euclidean distance metric gave a very high prediction of perceptual configuration.

On the basis of these preliminary analyses, some suggestions for further perception, auditory, and production, experiments are made.

## 2 Experiment 1: Perceptual analyses of 'natural' fricatives

### 2.0 Introduction

The purpose of this experiment was to establish a perceptual space for natural, syllable-initial fricatives. By 'natural' fricatives, we mean that the speech signals contain not only the fricative section, but also the following transitions and vowels. The effects of other syllabic positions on the perception of the fricatives, or the various vowel contextual effects, were of no concern here.

### 2.1 Stimuli

Clearly it would be desirable to include as many stimuli as is practically possible for any MDS analysis. This is because, with an increase in the number of stimuli, the number of dimensions which can be explored also increases. "Ideally", 12 stimuli are recommended for two-dimensional solutions and 18 stimuli for three-dimensional solutions (Schiffman *et al.*, 1981). But some studies recommend fewer stimuli than this. For example, Kruskal & Wish (1978) recommended 9 stimuli for two dimensions, 13 for three, and 17 for four. Schiffman *et al.* (1981) mentioned three possible conditions under which these recommendations can be violated: i) time constraints for collecting data, ii) unavailability of sufficient stimuli, or iii) to illustrate a specific point. Also, they observed that the recommended stimuli numbers are for a single matrix of data, and if the data sets have

several matrices, "the recommendations can be weakened somewhat".

In this experiment, the data set contained seven stimuli. The actual stimuli were natural productions of voiceless fricatives, [f, θ, s, ʃ, ç, x, h], followed by the vowel [ɑ]. The non-English syllable-initial fricatives [ç, x] were included to increase the number of combinations for the similarity judgements. According to the Pols *et al.* study (1969), the psychological vowel spaces of individuals who speak different languages show similar perceptual dimensions. They tested the same synthetic Dutch vowel stimuli set on two foreigners, one a Welshman, the other a native speaker of Japanese. Their three-dimensional vowel perceptual spaces closely correlated with those of native Dutch speakers. We are aware that their stimuli were wholly synthetic. Therefore, in this experiment, the acceptability of these two *natural* non-native fricatives was tested out, for a group of native English speakers; and if the results were consistent among the subjects, then they may be included in further experiments.

One trained female phonetician recorded these consonants with a repetition. The recording was done in an anechoic chamber onto a Sony DTC-1000ES digital audio tape recorder. They were digitised with a 16-bit quantization rate and a sampling rate of 20 kHz. The acquired materials were labelled and checked for any abnormality. The vowel sections were checked so that the peak intensity of each stimulus was kept within 2~3 dB across consonants difference. An example of the stimuli is given in Figure III-1.

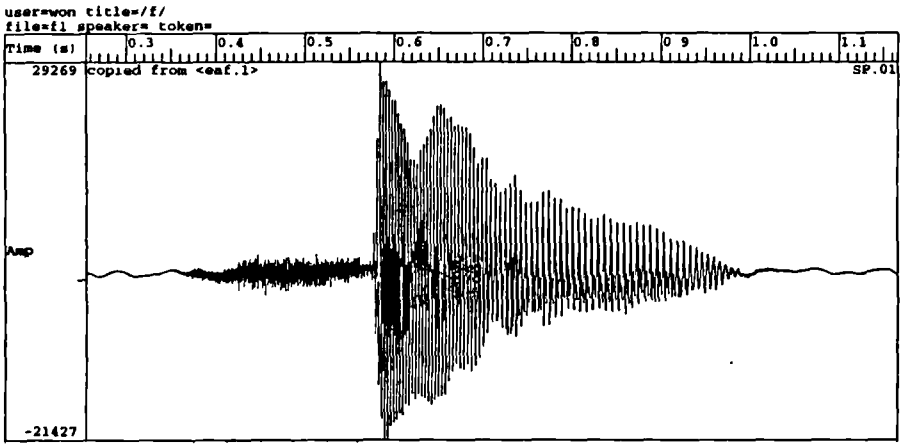



**Figure III-1.** An example of stimuli for experiment 1.

## 2.2 Data collection method

The most fundamental method of gathering perceptual data for MDS is a similarity

judgement, as stated by Schiffman *et al.* (1981):

> The *similarity judgements* are the *primary* means for recovering the underlying structure of relationships among a group of stimuli. It should be emphasized that the spatial arrangement derived by MDS from similarity judgements is the heart of the analysis. (p19)

An argument against the similarity judgement is that it is something of an artificial task, compared with an identification-type experiment. In an identification task, subjects simply listen to stimuli and identify them. Responses are entered in the appropriate cells in a stimulus-by-response matrix (= confusion matrix). However, this is a minor disadvantage compared to the inaccuracies involved in symmetrisation to fill up the zero confusions in the responses of uncorrupted stimuli, as we mentioned in the review of Klein *et al.* (§II.2.1.2).

There are many ways of extracting similarity judgements, including magnitude scaling, magnitude estimation, and triadic comparisons. For magnitude scaling, listeners are presented with a pair of sounds and required to give similarity judgements on a fixed scale of, say, 10 discrete points: the value '1' means the most similar sounding pair and '10' means the most dissimilar sounding pair. The intermediate numbers represent intermediate degrees of similarity. The psychological effect of the scaling is that listeners tend to rate the stimuli pairs around the mid values rather than the extremes. This problem can be overcome with extensive training, or subjects can be guided by a reference point. For example, subjects can be told the difference between [f] and [h] is 10 and ask them to base their judgements on this reference, but then, these references themselves may interfere with people's judgments.

In magnitude estimation, listeners judge the similarity of each pair of sounds in terms of a measurable distance on a line[1] which reflects the perceived distance between the sounds; longer lines represent more dissimilar pairs, while shorter lines represent more similar pairs (Singh, Woods & Becker, 1972). The scale of length can be chosen freely by each individual. With this method, however, It is necessary to make an unqualified assumption that there is a one-to-one relationship between auditory and visual sensation.

The triadic comparison task is a kind of extended paired similarity test. Three sounds are presented, and the subjects are required to judge whether the second or third

[1]Numbers can also be used.

sound was more similar to the first. This is the simplest comparison task a listener can perform, no response as to degree of similarity is required. With the extensive data collection time required for MDS analysis, this technique has an obvious advantage over other similarity judgement tasks, especially for stimuli with no acoustic corruption. Also, because the task requires very few instructions and little training, the collected data have very little experimental corruption.

Therefore, in this experiment, the specific task set for listeners was triadic comparison judgements; on each experimental trial, three non-identical stimuli were selected and paired into two groups for presentation, from the set of basic stimuli. That is, the stimuli within the triads ABC were paired to AB AC in order to help the short term memory of the listeners. 105 (=7×6×5/2) triads were constructed from these noises and randomised. An interval of 0.1 seconds between the tokens and 2 seconds endpause were used. The stimuli were generated and recorded onto a DAT tape with a pause and a tone every block of 5 stimuli for presentation to the listeners.

### 2.3 Subjects and procedure

Students studying M.Sc. in Speech and Hearing Sciences volunteered to listen to the stimuli. Since the test was conducted at the very beginning of the M.Sc. course, students were phonetically naïve listeners. All six subjects were native speakers of English, and had normal hearing. All but one of the subjects completed the test. They had no knowledge of the purpose of this study, nor of the experiment.

The test was carried out in a quiet room and subjects were given both a brief verbal introduction and the following written instruction:

---

In this experiment, we are interested in your immediate reactions as to how speech sounds differ.

You are going to listen to some sounds on the tape recorder and tick a relevant box for the more similar sounding pair. This task is, of course, subjective. There are no right or wrong answers. We just want to know which sounds are closer to each other to you.

Before we begin the actual test, we would like you to be familiar with the test items. All of the sounds to be used in the actual test are given at the beginning. Just listen to these sounds as each sound is presented.

Next there are five examples in order for you to familiarise yourselves with

the test method. The sounds will be presented in pairs of type AB AC, so that the first and the third sounds are the same. You are to listen carefully and give your judgement on the answer sheet by ticking the first or the second box corresponding to the first or the second pair of sounds you think are more similar or close to each other.

In the main test, you are asked to do the same process 105 times. After every block of five, there is a short pause and a beep so that you can check you are following the test.

Data were accumulated over trials according to the following scoring procedure: the pairs selected as more similar were assigned 1 scores, and the pairs which were not selected, 0 scores. In this way, a matrix of data indexing the perceived relationships among the seven fricatives was obtained for each subject. As an example, the matrix for a subject who rated the fricatives is shown in Table III-1.

| | f | θ | s | ʃ | ç | x | h |
|---|---|---|---|---|---|---|---|
| f | 0 | 5 | 3 | 4 | 2 | 1 | 0 |
| θ | 5 | 0 | 4 | 3 | 2 | 1 | 0 |
| s | 1 | 2 | 0 | 5 | 4 | 3 | 0 |
| ʃ | 2 | 3 | 4 | 0 | 4 | 2 | 0 |
| ç | 0 | 1 | 3 | 3 | 0 | 4 | 4 |
| x | 0 | 1 | 3 | 4 | 5 | 0 | 2 |
| h | 0 | 1 | 2 | 3 | 4 | 5 | 0 |

**Table III-1.** Similarities matrix for a subject who rated the fricative syllables.

Typically, the responses for pairs of type AB and BA were not symmetrical, as in the above matrix. Asymmetry of the responses is a common phenomenon for reverse ordered pairs, and the similarity matrix was normally symmetrised by simply adding the similarity responses between a pair of consonants $C_i$ and $C_j$ as below:

| | f | θ | s | ʃ | ç | x | h |
|---|---|---|---|---|---|---|---|
| f | | | | | | | |
| θ | 10 | | | | | | |
| s | 4 | 6 | | | | | |
| ʃ | 6 | 6 | 9 | | | | |
| ç | 2 | 3 | 5 | 7 | | | |
| x | 1 | 2 | 6 | 6 | 9 | | |
| h | 0 | 1 | 2 | 3 | 8 | 7 | |

**Table III-2.** Symmetrised matrix of Table III-1.

## 2.4 Analyses

The symmetrised similarity matrices obtained from each of the listeners were then examined using MDS analyses. This technique, as mentioned before in §I.3.3, was used to account for the perceived differences between pairs of stimuli by locating the stimuli in an underlying n-dimensional perceptual space. There were no *a priori* assumptions made concerning the number or the nature of the dimensions underlying the listeners' perception of the stimuli.

We have mentioned that in speech perception experiments the MDS is performed not on one set of distance measurements but on a number of them. The number of measurements corresponds usually to the subject number involved in the perceptual test. This MDS technique, which takes into account subject variability into the spatial configuration of stimuli, is called INDSCAL. Confusing enough, the original individual difference scaling program developed by Carroll & Chang (1970) was also called INDSCAL and so were many versions modelled after this. More accurately, this MDS was called **Weighted MDS**[1] (henceforth WMDS) or 3-way MDS, since the third way of the data may be occasions or experimental conditions, as well as individuals (as in Soli & Arabie, 1979; Baker & Rosen, 1993). The basic concepts of WMDS model are illustrated below with a hypothetical example.

WMDS produces two spaces — **group stimulus space** which shows the perceptual organisation of stimuli for all subjects, and the **subject space** which indicates the measure of individual subject differences. How the subject space can be interpreted is best illustrated by a simple example. Suppose we have a stimulus space with nine stimuli arranged in a square, as shown in Figure III-2 (a). And the corresponding subject space is shown in Figure III-2 (b), which indicates the **weights** or perceptual salience of the stimulus dimensions for nine hypothetical subjects.

[1]Schiffman *et al.* (1981) use this terminology.

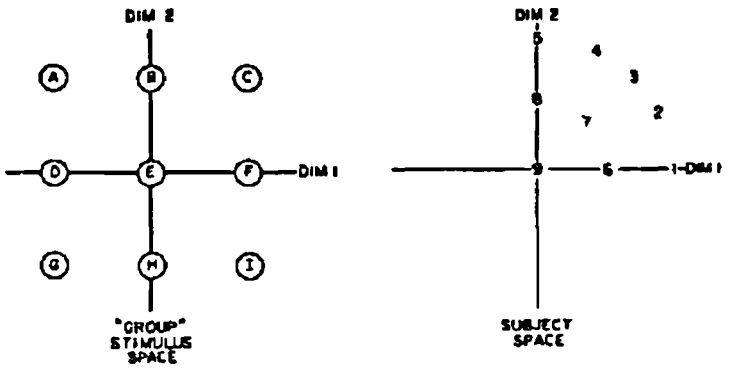



**Figure III-2.** (a) A hypothetical group stimulus space from a 3-way MDS analysis, (b) A corresponding subject space; after Schiffman *et al.* (1981).

These weights can be thought of as stretching or contracting factors to the dimensions of the group stimulus space. Thus the individual subject's perceptual space can be obtained by multiplying the weights to the group space. For example, the Subject 3 places equal weights on both dimensions. This means that Subject 3's private perceptual space looks exactly the same as the group stimulus space. However for Subjects 2 and 4, their private perceptual spaces look like Figures III-2 (c) and (d). Subject 2, who gives more emphasis on dimension 1 than 2, has a stimulus space stretched along the dimension 1 axis. The reverse effect is shown for Subject 4, who gives more weight on dimension 2 than 1; that its stimulus space is stretched along the dimension 2 axis.

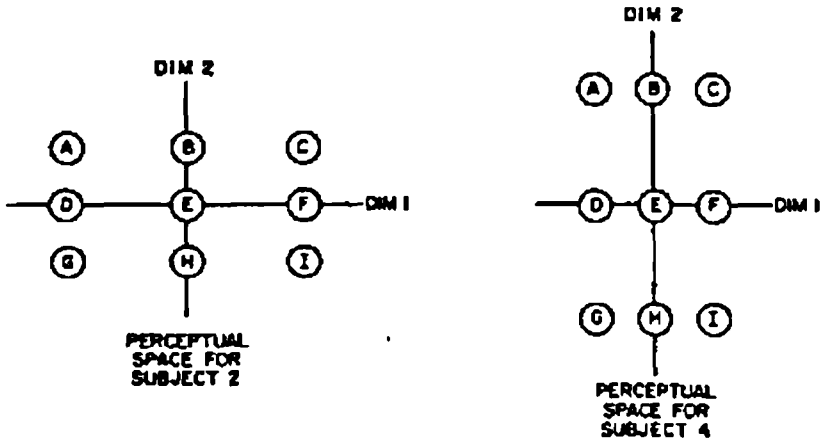



**Figure III-2.** Individual perceptual spaces for (c) Subject 2, (d) Subject 4; after Schiffman *et al.*

Distance of a subject from the origin in the subject space indicates how much of the variance of that subject's data is accounted for by the particular MDS solution (the stimulus space). The Subjects 3 and 7 have the same pattern of dimension weights, but the MDS solution reflects the Subject 3's data more accurately than 7's. Thus, data for subjects who are closer to origin are less well accounted for by the MDS solution. The given group stimulus space is totally unsuitable for Subject 9, and other dimensions may be necessary to explain this subject's data, other than the first two dimensional solution shown here.

The actual MDS program used for the analyses was a version of INDSCAL (a program written by Bosman, University of Amsterdam). This programme consists of two parts: **metric (interval)** scaling, as proposed by Carroll & Chang (1970), and an intermediate phase using **nonmetric (ordinal)** scaling, similar to the one by Kruskal (1964). The latter was incorporated because it is "very effective in increasing the convergence". Since the final step in each iteration consists of the metric part, the result is a metric scaling. Like any other MDS programme, the object of the scaling is to obtain an optimal spatial representation of the scaled objects (fricatives), representing them in several interpretable perceptual spaces.

Later, the matrices were also analysed using proc-ALSCAL program (SAS windows version 6), to try out the nonmetric solution. By comparing metric and nonmetric solutions, we can check the stability of the configurations[2].

## 2.5 Results

### 2.5.1 Metric analyses

The choice of the optimal INDSCAL solution was based on a consideration of the amount of variance accounted for by a various number of dimensional solutions, and the interpretability. Figure III-3 presents fit curves showing the cumulative variance accounted for as a function of the number of perceptual dimensions.

[2]In Chapter IV, both metric and nonmetric MDS are carried out by proc-ALSCAL program. But in the preliminary analyses, nonmetric solution was obtained by INDSCAL. For details, see §IV.5.

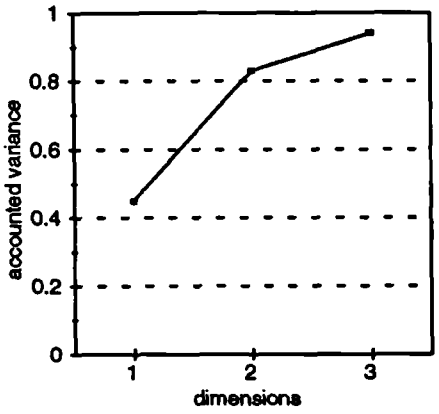



**Figure III-3.** Fit curve representing the cumulative variance accounted for by metric (INDSCAL) multidimensional scaling analyses as a function of the number of dimensions.

The INDSCAL programme used here did not permit the determination of one- or four-dimensional solutions, but the fit curve shown for 3-dimensional solution demonstrates that the cumulative variance accounted for becomes very gradual after two dimensions. This is commonly known as an "elbow". That is, most of the variance was accounted for by the first two dimensions (0.83); adding a third dimension accounted for an additional 0.11 of the variance in the listeners' perceptual data. Thus, the two-dimensional space was chosen as the solution of the INDSCAL analysis, but the three-dimensional space will be also presented here for a comparison, with the results of later nonmetric analyses.

The perceptual map of the two-dimensional solution is shown in Figure III-4. The most noticeable feature of the map is that the fricatives form three separate groups, [f θ], [s ʃ], [ç x h]. Thus, Dimension 1 (horizontal-axis) can be given a 'place of articulation' label. On Dimension 2 (vertical-axis), it can be said that the sibilants [s, ʃ] are separated from the non-sibilants [f θ ç x h], but again, what is more prominent is the fact that the fricatives are arranged into three distinct groups. This is what we would expect if the subjects were using a phonetic criterion rather than a psychoacoustic one when they are making judgments.

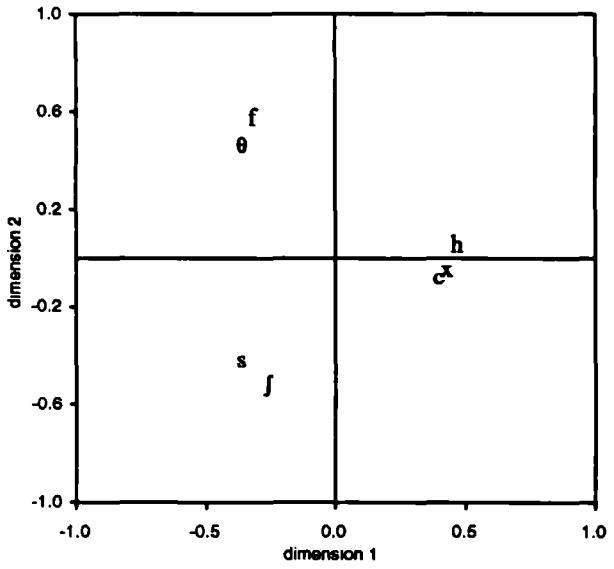



**Figure III-4.** The two-dimensional INDSCAL solution obtained for 5 listeners.

The INDSCAL procedure also lists 'individual weights' for each dimension, which gives an indication of how much perceptual importance the subjects placed on each dimension. In Figure III-5, the subject weights of dimension 2 were plotted against the weights of dimension 1. This indicates, for example, that the similarity judgements of subjects 1 and 4 were dominated by dimension 1, whereas the judgements of subjects 2 and 3 put more importance on dimension 2. Subject 5 gave almost equal perceptual weights to the two dimensions. Although there seems to be considerable variability between the subjects, what is most significant in interpreting the subject data in the INDSCAL analysis is the distance of a subject from the origin in the subject space. This distance is roughly proportional to the variance of the data for a particular subject accounted for by the multidimensional solution (§2.4). Thus, all of the five subjects' data are well accounted for by the stimulus space.

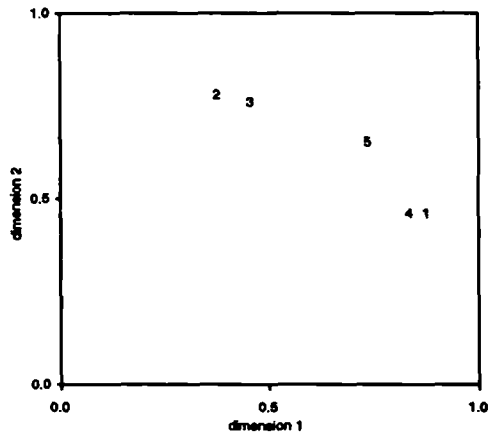



**Figure III-5.** Subject weights for dimensions 1 and 2.

The three-dimensional solution of the result is shown in Figure III-6. Dimensions 1 and 2 are almost identical to those of the two-dimensional solution. Dimension 3 separates the fricatives which were closer together in dimension 2. Neither phonetic nor acoustic interpretation is possible. This confirms that the two-dimensional solution is appropriate for the data.

(a) (b)

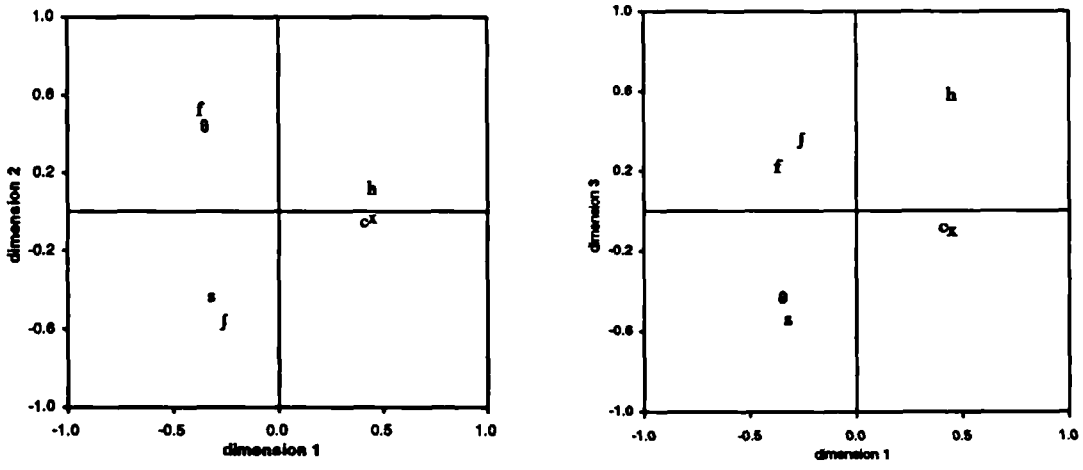



**Figure III-6.** The 3-dimensional INDSCAL solution: (a) dimension 2 against dimension 1; (b) dimension 3 against dimension 1.

### 2.5.2 Nonmetric analyses

SAS/STAT for version 6 is used for the nonmetric analyses. The analyses were carried out iteratively, with random starting configurations, feeding the output of the previous result as a starting point for the next analysis, to ensure a global minimum is reached. Figure III-7 shows a plot of the badness-of-fit against the number of modelling dimensions of one to five. Increasing the number of dimensions reduces the error of the fit, as expected. The error is reduced exponentially; there is a relatively large reduction in badness-of-fit for the shift from one to two dimensions, a smaller one for the shift from 2 to 3, and smaller still for 3 to 4 dimensions.

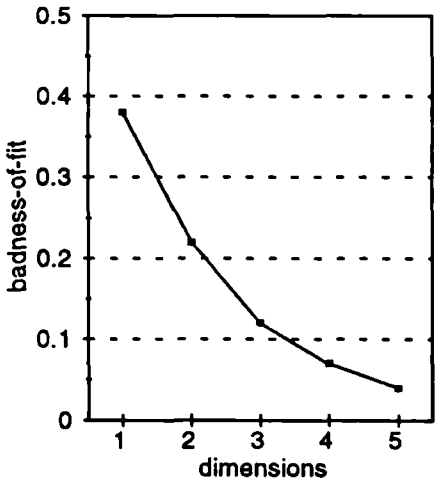



**Figure III-7.** Badness-of-fit curve representing the fit error by nonmetric (proc-ALSCAL) multidimensional scaling analyses as a function of the number of dimensions.

On the basis of these statistical data, it was difficult to decide upon the dimensionality of the stimulus set; two, three or four dimensions may be equally appropriate. So each dimension had to be considered in terms of the interpretability. Thus, stimulus configurations for two, three and four dimensional solutions are plotted in Figures III-8 to III-10, to compare the interpretability of the each dimension.

Figure III-7 shows dimensions 1 and 2 of the two-dimensional solution. Unlike the two-dimensional solution of the INDSCAL analysis, the place of articulation dimension is not clearly shown. Also, the fricatives are more scattered on the space.

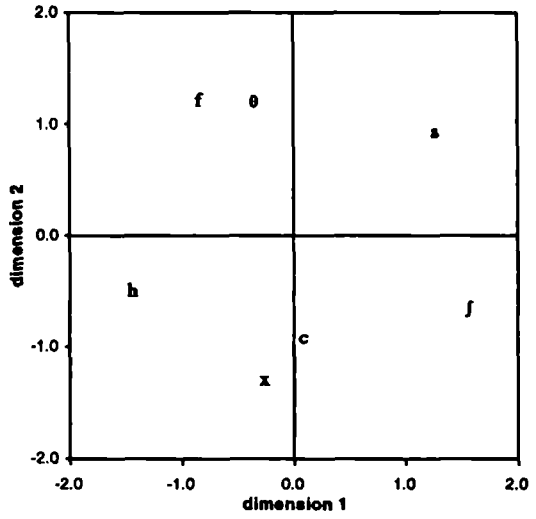



**Figure III-8.** The 2-dimensional nonmetric solution.

Figure III-9(a) shows dimensions 1 and 2 of the three-dimensional solution. The configuration shows a strong resemblance to the metric solution. That is, the fricatives form three groups according to their place of articulation. Figure III-9(b) is a plot of dimension 3 against dimension 1. The configuration gives a very scattered look, but in fact, what dimension 3 has done is to move apart the fricatives which were close together in dimension 2, except for the fricative, [ç]. The latter seems not to be grouped with the other two back fricatives in dimension 2, but seems to lie independently.

(a) (b)

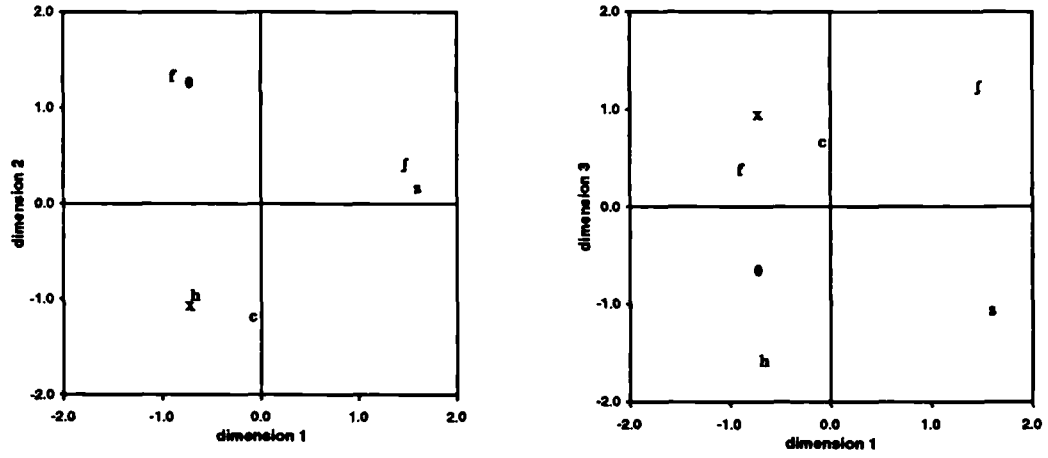



**Figure III-9.** The 3-dimensional solution of the nonmetric analyses.

The weighting of individual subjects for a three dimensional solution is given in Table III-3. For all three dimensions, most of the weights are of similar value, except for subjects 3 and 5, indicating that these dimensions are of equal importance for all individuals. Less emphasis is given to dimension 3 for subject 4 and dimension 2 for subject 5.

| subjects | dimension weights | | |
|---|---|---|---|
| | dimension 1 | dimension 2 | dimension 3 |
| 1 | 0.98 | 1.02 | 1.00 |
| 2 | 1.03 | 1.00 | 0.97 |
| 3 | 0.95 | 1.07 | 0.98 |
| 4 | 1.15 | 1.25 | 0.33 |
| 5 | 1.13 | 0.61 | 1.17 |

**Table III-3.** Dimension weights for the 3-dimensional model for each of the 5 subjects.

Figure III-10 shows the configurations for the four-dimensional solution. The plot of dimensions 1 and 2, in Figure III-10(a), shows an even stronger grouping of the fricatives according to their place of articulation. Figure III-10(b) is the plot of dimension 4 against dimension 3. This shows that on these two dimensions, the fricatives which were close together in dimensions 1 and 2 are moved apart; note that fricative pairs [f θ], [s ʃ] and [x h] are those which are moved diagonally apart.

In summary, two-dimensional solution[3] is adequate to represent the perceptual data. It is clear that the subjects were using a phonetic measure of similarity and the two MDS dimensions correspond to the 'place' and 'sibilance' properties of fricatives. In particular, the placement of fricatives on the place dimension seems to reflect a decreasing anterior cavity size as place of constriction moves from back to front (in Figure III-4).

[3]For the nonmetric analysis, the first two dimensions of three-dimensional solution were only interpretable.

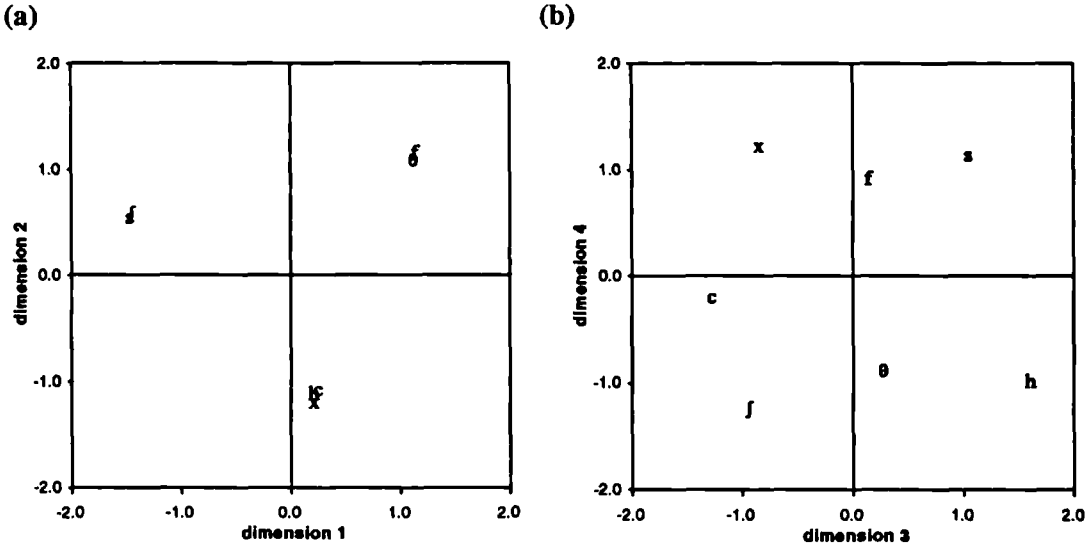



**Figure III-10.** The 4-dimensional nonmetric solution.

## 2.6 Discussion

The aim of this experiment was to make a preliminary investigation of the perceptual space of the natural fricatives. But, given the diversity of the MDS analyses that is possible for the set of perceptual data presented above, a comparison of the metric versus nonmetric analyses is pertinent. The perceptual effect of those non-English fricatives included in the experiment also needs also to be discussed.

- The proc-ALSCAL analyses using the SAS system provide, in general terms, a much more versatile and reliable tool for conducting the MDS process. It allows the user to explore all the different combinations of MDS techniques. For example, switches between metric/non-metric, 2-way/3-way, triangular/square (symmetric/asymmetric) input matrices, and many more besides, are all possible. However, all the manipulations have to be done by a SAS specific language. This can be an extremely lengthy process. In contrast, the INDSCAL is an easy-to-use programme with built-in graphics. This was why the metric analyses using the INDSCAL program was first carried out to examine the results. However, the INDSCAL program has its limitations. For example, Proc ALSCAL was capable of analysing many different dimensional solutions appropriate to the data here, whereas INDSCAL could only cope with two- and three-dimensional solutions.

However, there is no clear reason why the configurations in the two-dimensional solution of the nonmetric analysis were different from those of dimensions 1 and 2 of the three-dimensional solution. This was in contrast to the configurations of dimensions 1 and 2 of the two- and three-dimensional solutions of the metric analysis, which were almost identical. The weighting for individual subjects in both metric and nonmetric solutions was comparable. Overall, the results from both analyses were the same for the particular stimulus set used here; the three- and four-dimensional solutions of nonmetric analyses and the solutions of metric analyses show the role of place of articulation as the most prominent in judging similarities between the fricative pairs. However, the nonmetric analyses may be slightly favoured because of the nature of the data collection method; it seems reasonable to view the similarity pair judgements examined here as ordinal-level data. On the other hand, ALSCAL at the ordinal level is reported to show the tendency to "compress the differences among both stimuli and subjects" (Schiffman, 1981: p238). In the next experiment, nonmetric ALSCAL analysis will be first performed, and the stability of this solution will be checked with the INDSCAL solution.

- The other point which has to be discussed here is the effect of the non-English fricatives [ç] and [x] on the perception of the stimulus set, which was made up of otherwise homogenous English fricatives. One subject could not complete the test and complained that stimuli pairs with [ç] had been especially confusing and at odds with the remaining pairs; these were described as having "different vowel quality". The other subjects also complained that the 'foreign' sounding consonants were a disturbing feature of the test. Along with the problem of subjects' confusions over non-English fricatives and the time consumed for including the fricatives [ç x], it was thought better to exclude these fricatives from further perceptual testings.

# 3 Experiment 2: Preliminary analyses on phonetic, perceptual and auditory spaces of fricative sounds

## 3.0 Introduction

Experiment 2 was designed to replicate the two experiments, which were described in detail in §II.2.1, by Pols and his colleagues (1969,70). As a brief reminder of the previous experiments, recall that, Klein *et al.* (1970) used 100 ms vowel segments extracted from the 12 Dutch vowels produced by 50 different speakers, for an identification test. The earlier work in 1969, by Pols *et al.*, used 405 ms vowel-like sounds, which were replicated single pitch periods that had been extracted from each of the 11 Dutch vowels spoken by one of the authors. The perceptual configurations from the Klein *et al.* experiment was clearer, in the 'linguistic' sense. More interestingly, though, the perceptual results of both experiments correlated highly with the corresponding acoustic data. Therefore, the stimuli in this experiment are divided into two different sets.

The purpose of this experiment was to observe the correlations between phonetic, perceptual, and auditory spaces, on a small set of fricative data. §3.1 is about the MDS analyses of the perceptual judgment, and possible phonetic interpretation of the perceptual space. §3.2 focuses on auditory analyses, in terms of critical band pass filtering and distance modelling. §3.3 questions the relationship between perceptual and auditory spaces. §3.4 discusses possible implications of the results, which lead to further experiments in subsequent chapters and to experiment 3.

## 3.1 Perceptual analyses

### 3.1.1 Stimuli

*Set 1.* Klein *et al.* extracted the stimuli from 50 different speakers. An experiment on such a scale, however, was not possible for present work. Thus, the fricative portions were cut out from the same stimuli used in experiment 1, ensuring that the vowel transition section was excluded. This process was carried out manually; the end of the fricative part and the beginning of the transition part were clearly visible from the narrow-band and the wide-band spectrographic analysis.

Loudness normalisation was not carried out.

*Set 2*. Since fricatives are not periodic sounds, pitch-period iteration technique was not suitable for synthesising the fricatives. Instead, the fricatives were synthesized by LPC autocorrelation technique with ten coefficients, 20 kHz sampling rate, from each fricated portion of materials used in set 1, so that the average spectral characteristics of natural speech were still retained. The length of each fricative was fixed at 0.5 seconds and the auditory loudness was normalised with a sound pressure level meter (A weighting)

The examples of resulting stimuli from each set are shown in Figure III-11.

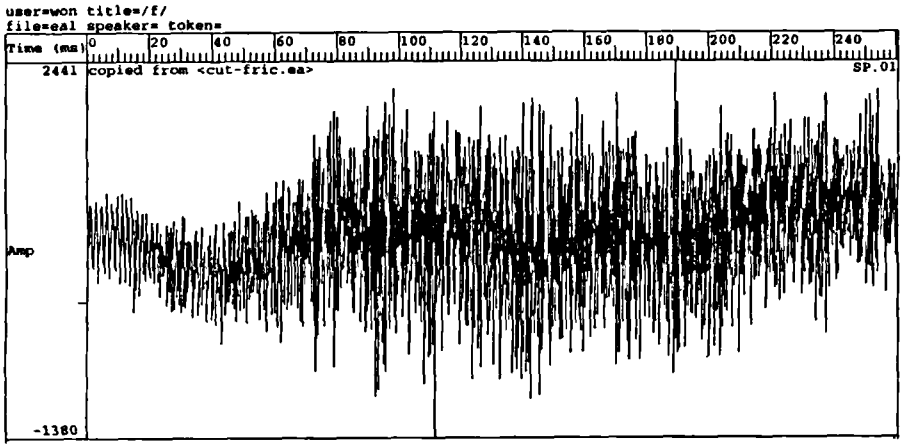



**Figure III-11.** (a) An example of natural fricatives without the transition part used for the stimuli set 1.

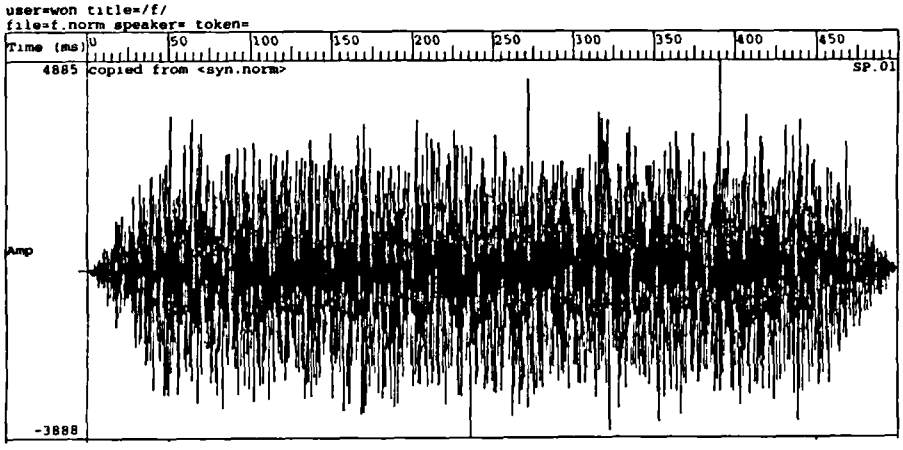



**Figure III-11.** (b) An example of LPC encoded fricatives used for the stimuli set 2.

The method for perceptual tests was the same as in experiment 1 ('pairwise similarity judgement').

### 3.1.2 Subjects

Ten undergraduates studying Speech Science were asked to participate in this experiment, five subjects for each stimulus set. All of them were native speakers of English, and none had any history of hearing difficulties. They had no prior knowledge of the purpose of this study or of the experiment.

### 3.1.3 Perceptual space

#### 3.1.3.1 Set 1

Nonmetric ALSCAL did not permit the determination of solutions of more than two dimensions for this data. Badness-of-fit for the one-dimensional solution was 0.29. For the two-dimensional solution, badness-of-fit decreased to 0.11. Figure III-12 is a plot of the one-dimensional solution.

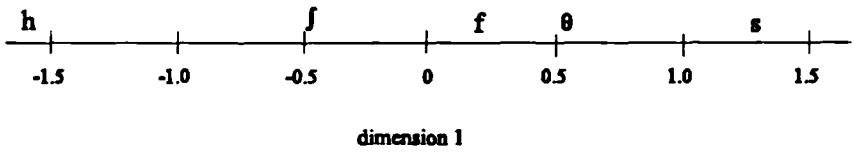



**Figure III-12.** The one-dimensional nonmetric solution.

The one-dimensional solution can not be interpreted for any known descriptions of the fricatives.

Figure III-13 shows the two-dimensional solution of the perceptual responses for the stimuli set 1. The fricatives are not grouped together as for experiment 1, but in dimension 2 it can be said that the fricatives were placed in the order of [f, θ, s, ʃ, h], corresponding to the place of articulation feature. Dimension 1 can be matched to the sibilance feature, but this is not so prominent.

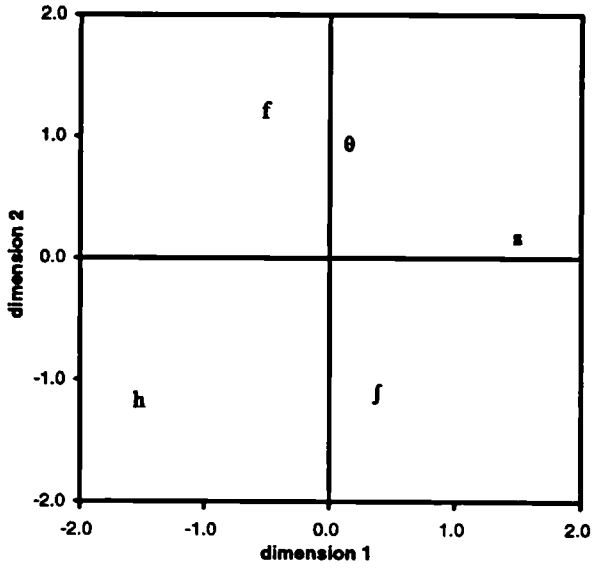



**Figure III-13.** The 2-dimensional nonmetric solution obtained from the stimulus set 1, the cut-out fricatives.

Subject weights of the two-dimensional solution is given in Table III-4. For subject 3, dimension 2 (the place dimension) is much less important than dimension 1 (sibilance dimension). The other subjects give similar weights to both dimensions.

| subjects | dimension weights | |
|---|---|---|
| | dimension 1 | dimension 2 |
| 1 | 1.13 | 0.85 |
| 2 | 0.91 | 1.08 |
| 3 | 1.30 | 0.56 |
| 4 | 0.84 | 1.14 |
| 5 | 1.16 | 0.81 |

**Table III-4.** Subject weights for the 2-dimensional solution for each of the 5 subjects.

The perceptual data were also analysed by INDSCAL, and the configurations of the 2-dimensional solution were almost identical to the nonmetric solution.

### 3.1.3.2 Set 2

For stimulus set 2, badness-of-fit for the one-dimensional solution was 0.3; for two-dimensions, it decreased to 0.15; for three-dimensions, it decreased to 0.04. Again, badness-of-fit values by themselves cannot determine the dimensionality of the analysis. So, both two-, and three-dimensional configurations are plotted.

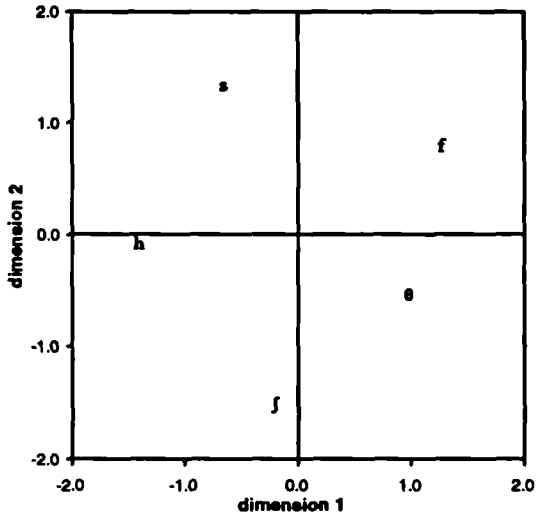



**Figure III-14.** The 2-dimensional nonmetric solution of the stimulus set 2, the LPC synthesised fricatives.

Dimension 1 of the two-dimensional solution (Figure III-14) can only tentatively be interpreted as a place of articulation dimension. Dimension 2 is not interpretable. This may be attributed to the loudness normalisation, that is, the fricatives [s ʃ] were no longer louder than the other fricatives, as in set 1.

The configurations for dimensions 1 and 2 of the three-dimensional solution is presented in Figure III-15(a). These were very similar to those of the two-dimensional solution. Plotting the third dimension does not improve interpretability (Figure III-15(b)).

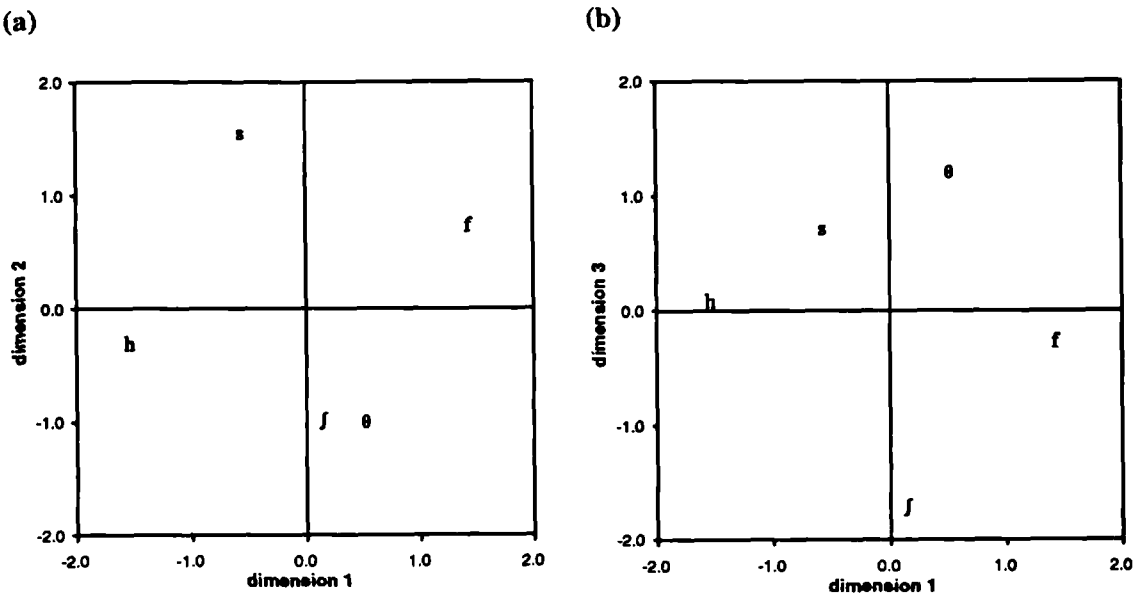



**Figure III-15.** The 3-dimensional nonmetric solution of the LPC synthesised fricatives.

Therefore, the two-dimensional analysis seems to be adequate here. The subject weights for dimensions 1 and 2 were very similar.

Some difference was observed between INDSCAL solutions and the nonmetric ones, but the difference was the same as that observed in experiment 1. That is, the configurations of the two-dimensional solution by the INDSCAL analyses were similar to the first two dimensional configurations of the three-dimensional solution in the nonmetric analyses. But, essentially the configurations were stable.

Having established the perceptual configurations, let us now examine the auditory configurations.

## 3.2 Auditory analyses

### 3.2.0 Introduction

In order to obtain the auditory space of the stimuli used in §3.1, three separate analyses were carried out. Firstly, the spectra were processed by 1/3 octave bandpass filtering to model frequency analyses in the auditory periphery. After this filtering, distances between the spectra were calculated by three different metrics — Euclidean, Slope, and N2D metrics. These spectral distances were finally transferred to MDS configurations on an auditory space. Details are given below.

### 3.2.1 Critical bandpass analyses: 1/3-octave bandpass filtering

Auditory analyses were based on the well-accepted concepts of critical band-pass filtering and logarithmic scaling of loudness. It could be argued that the spectral distances calculated using spectra which include a more detailed auditory masking pattern should be included. It is reported that the dominant frequencies model, which takes such properties of the peripheral auditory system into consideration, gives a markedly better result for the phonetic judgement of vowels (Carlson & Granstrom, 1979). This claim is supported in the synthetic vowel study by Bladon & Lindblom (1981) and the synthetic consonant study by Sidwell & Summerfield (1986). In both cases, spectral shape was derived from the auditory masking filter model or simultaneous-masking technique.

The positive findings of the effect of masker filtering in the perception of synthetic stimuli have been, however, contradicted in the experiments using natural speech. Using Swedish vowels and consonants, Blomberg *et al.* (1986) showed that, for their recognition system, the best predictions were obtained with simple dB versus a linear frequency scale (97%) without the masking filters, whereas the average recognition accuracy for the dominant frequencies model was 94% (99% for vowels; 90% for consonants). Furthermore, it is speculated (Klatt, 1982b) that,

> dominant frequencies processing may simply be the way that the auditory system solves the problem of producing a spectral representation whose shape is more-or-less insensitive to input level, and engineering systems that do not have the severe dynamic range problem seen in the average firing rate behaviour of a primary auditory neuron need not go to the trouble of computing dominant frequencies" (p194).

It seems that the perceptual consequences of masking in auditory representation are inconsistent, and that further work is required. Thus, for the purposes of our investigation, a critical band and logarithmic scaling of loudness model are assumed to be adequate as a first stage of the peripheral auditory analysis.

The method of 1/3-octave band filters, used in Pols *et al.* (1969), is extended to the acoustic analysis of the fricatives here. This is because it can be expressed by a relatively simple formulation to reflect the critical bandwidth of the peripheral auditory system, over a large frequency range. The Mel frequency, Bark or ERB scales could have

been utilised equally well. In the auditory filter program ('afbanklist', written by Mark Huckvale, UCL), a 1/3-octave bandwidth was formulated as:

fwidth := exp(ln(2.0)/3)

This is the cube root of 2, which is about 1.26. So the bandwidth increases progressively by, 100, 126, 159, 200, 252, 317, 400, 505,..., from filter to filter.

The frequency bands range from 100 to 10,000 Hz, increasing in steps of:

fstep := exp ( ln(10000)-ln(100) )/ number of filters

If the number of filters is 32, the centre frequency will increase from 115 Hz to 132, 152, 175, 201, ..., till the centre frequency reaches 10,000 Hz.

A plot of this auditory filter bank is shown in Figure III-16.

The output of this program is the spectral energy levels measured in decibels for a specified number of frequency channels, on a 10 ms frame rate.

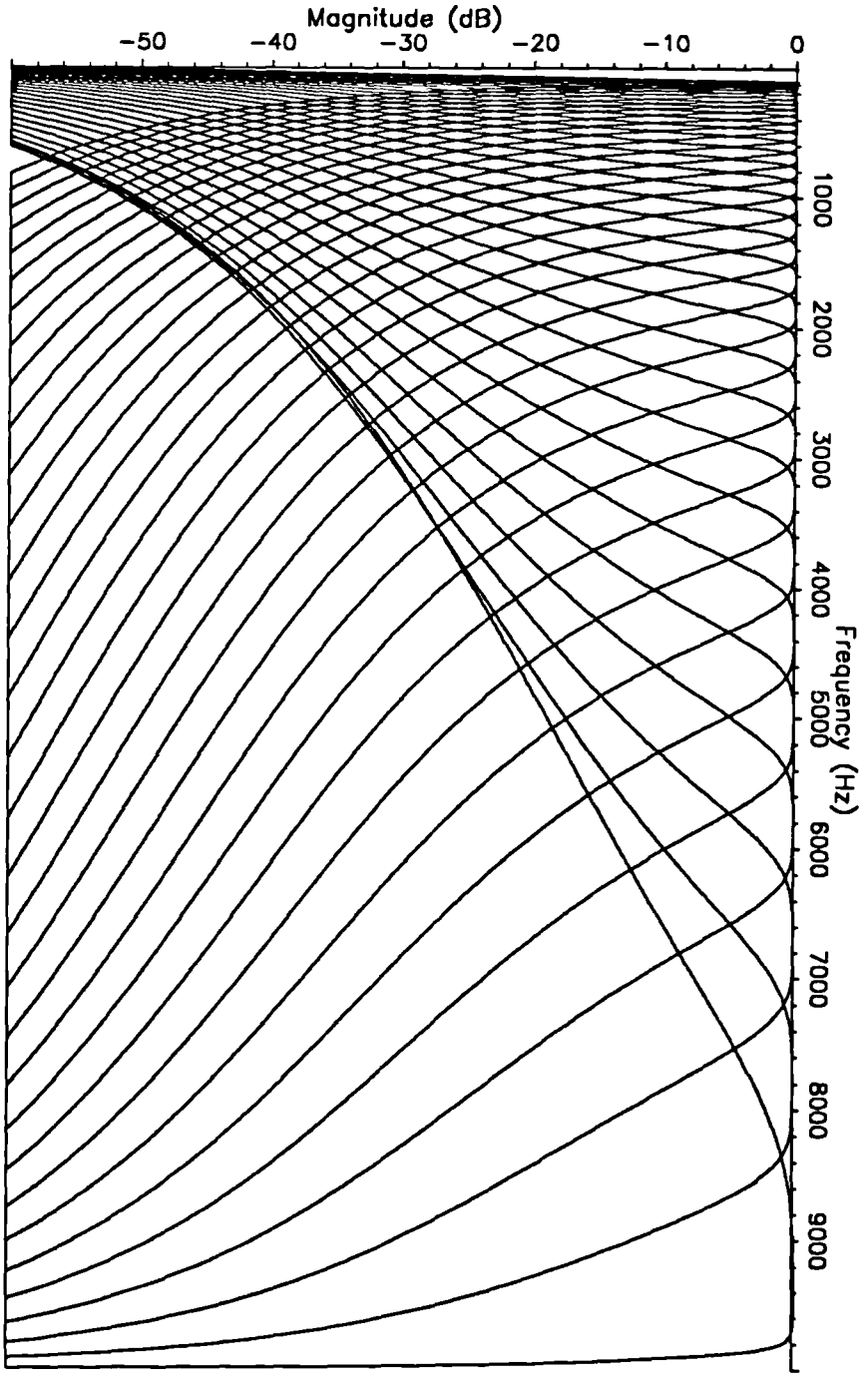



**Figure III-16.** A plot of auditory filter bank; 32 channel 1/3-octave filters.

### 3.2.2 Non-linear time alignment

To account for the differences in the length of the segments, the average spectral distance is calculated after a *non-linear time alignment* between the two segments is found. The process of finding non-linear alignment is best illustrated with a simple diagram as in Figure III-17.

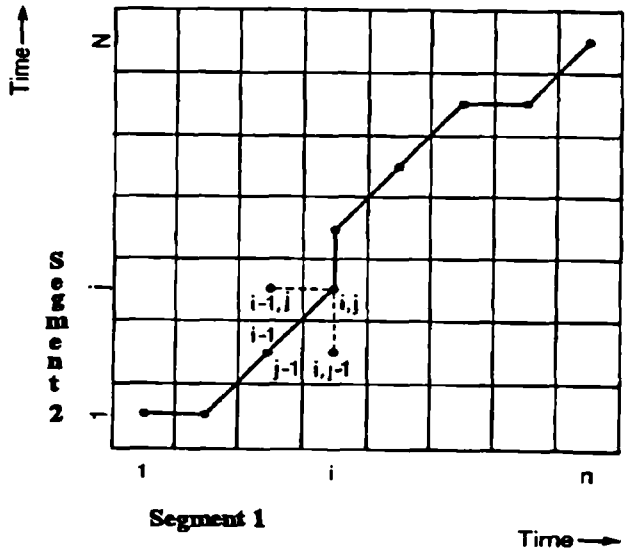



**Figure III-17.** Illustration of a time-alignment path between two segments that differ in time scale. Any point i, j can have predecessors as shown; after Holmes (1988).

Suppose that the two segments to be compared have n and N number of frames. The squared difference $d(i,j)$, between frame $i$ of segment 1 and frame $j$ of the segment 2, is calculated for each frame. To find the time aligned differences between the two segments, we must find the sum of the squared differences between the individual pairs of frames, along whichever path between the bottom left and top right corners in Figure III-17 gives the smallest distance. This process is done with a so-called dynamic programming algorithm (Sakoe & Chiba, 1978).

### 3.2.3 Distance Metrics

As mentioned in §II.2.1, Pols *et al.* (1969) used the sound pressure levels in dB in 18 filter bands for each of the 11 vowels, which constituted an 11 × 18 data matrix. In a geometrical model, the sound spectra of the 11 vowels result in a set of 11 points in an

18-dimensional space. By a principal-component analysis, the spectral relationship was represented in a three-dimensional solution. Distances between the vowels were taken as the length of the line joining the vowels in the 3-dimensional space. Thus, the acoustic vowel distance between two spectra *i* and *j*, can be expressed as:

$$ED = ( \Sigma (E_{i,k} - E_{j,k})^2 )^{1/2}$$

where $E_{i,k}$ is the energy in decibels in the *k*th filter of the spectra *i*. This is the square root of the squared differences, and is known as the Euclidean distance. However it must be noted that, although spectral peaks are known to have more perceptual weight than troughs, the Euclidean metric gives equal weight to peaks and troughs.

For a comparison of two excitation patterns which have the same peak locations but varying slopes of shoulders around the peaks, the Euclidean metric has been considered to be unsuitable (Klatt, 1982a). As the difference between the slopes increases, the distance calculated from Euclidean metric would increase, whereas the perceptual distance would remain unchanged. This was the result of the perceptual analysis by Klatt (1982a), who suggested the Weighted Slope Metric (WSM) which emphasises the formant frequency values but is insensitive to relative formant amplitudes, and to spectral tilt changes. The WSM distance between two spectra, $S_1$ and $S_2$, with N channel filters is given by:

$$WSM = k_E |E_{S1} - E_{S2}| + \Sigma 0.5 [k_{S1}(i) + k_{S2}(i)] [S'_1(i) - S'_2(i)]^2$$

where $E_{S1}$ and $E_{S2}$ are the overall energy levels and $S'_1$ and $S'_2$ are the spectral slopes given by the first difference:

$$S'_1(i) = S_1(i+1) - S_1(i), \qquad \text{for } i = 1, ..., N-1$$

The weighting function $k_S$ for the *i*th channel is given by:

$$k_S(i) = \frac{k_{Lmax}}{[k_{Lmax} + (E_i - E_{Lmax})]} * \frac{k_{Gmax}}{[k_{Gmax}(E_{Gmax} - E_i)]}$$

where $E_{Gmax}$ is the global maximum and $E_{Lmax}$ is the nearest local spectral maximum to S(*i*). It was explained that this coefficient was used for proportionate distribution of the

spectral weighting among the local and global spectral characteristics. At the spectral peaks, the constant $k_S$ has small values because the log spectral differences, $E_i - E_{Lmax}$ and $E_{Gmax} - E_i$ have similarly small values. Accordingly, the spectral slope difference, $S'_1(i) - S'_2(i)$, is emphasized at the spectral peaks.

It must be noted that the constants, $k_E$, $k_{Gmax}$, $k_{Lmax}$ were left unspecified and needs to be set by users. This means that, when $k_E$ is assigned a large value, the metric will be more sensitive to overall level differences. When the values of $k_{Gmax}$ and $k_{Lmax}$ decrease, sensitivity to differences between slopes near local and global maxima is increased. Very high values of $k_{Gmax}$ and $k_{Lmax}$ make the metric insensitive to the local and global peaks. In a particular case, when $k_E$ is set to zero and $k_{Gmax}$ and $k_{Lmax}$ are assigned to $\infty$, WSM metric approaches a Euclidean distance between the slopes, $S'_1$ and $S'_2$.

Nocerino *et al.* (1985) tested the performance of six different spectral distortion measures on isolated words speech recognition tasks and reported that the WSM analysis, on 23-channel spectral-band spectra, with the consonant specification, $k_E = 0$, $k_{Gmax} = k_{Lmax} = \infty$, resulted in an optimal correlation.

Assmann & Summerfield (1989) tested other metrics for optimal prediction of double vowel perception. The weighted negative second differential metric (WN2DM) by Assmann & Summerfield (1989) constitutes a slight modification of the WSM; the weighted function was retained but the slope term, $[S'_1(i) - S'_2(i)]^2$, is replaced by different representations.

The WN2DM, in place of $S'_1$ and $S'_2$, takes the absolute value of the negative part of the second differential of the excitation pattern, where $S''_1$ is computed as:

$$S''_1(i) = \max \{ - [ S_1(i - 1) - 2S_1(i) + S_1(i + 1) ], 0\}$$

The effect of this metric can be illustrated by considering a sine function, say, $y = \sin(x)$. The second differential of this function is a negative sine function, $y = -\sin(x)$. The negative second differential function then sets the positive portion to zero and takes the absolute value of the negative portion. Now, comparing with the original sine function, the peaks are preserved while the influence of the troughs becomes zero.

Table III-5 summarises previous distance metric studies, with their correlation values between the perceptual and acoustic distance data and their optimisation constants.

| | metrics | stimuli | unweighted | optimised | no. of filters & optimal weights |
|---|---|---|---|---|---|
| Pols *et al.* (1969) | EM | synthetic vowels | [1] 0.992, 0.971, 0.742 | | 18-channels |
| Kewley-Port & Atal (1989) | EM (Bark scale) | synthetic vowels | 0.93 | | |
| Klatt (1982) | EM WSM | synthetic /ɑ/ vowels | 0.85 | 0.93 | 36-channels $k_E$ =0, $k_{Gmax}$ = 20, $k_{Lmax}$ = 1 |
| Nocerino *et al.* (1985) | WSM | isolated words | [2] 8.45 | | 23-channels |
| Assmann & Summerfield (1989) | WLM | double vowels (synthetic) | 0.74 | 0.81 | 125-channels, $k_E$ = 0.000193, $k_{Gmax}$ = 1.5687, $k_{Lmax}$ = 106136.4 |
| | WSM | | 0.89 | 0.93 | $k_E$ = 0, $k_{Gmax}$ = 1.0546, $k_{Lmax}$ = 633182 |
| | WN2DM | | 0.92 | 0.94 | $k_E$ = 0, $k_{Gmax}$ = 6.3813, $k_{Lmax}$ = 70197855.4 |

[1] Correlation values for the first three physical and perceptual dimensions. [2] Average recognition error of the four talkers.

**Table III-5.** A summary of previous distance metrics studies.

It must be noted that each metric was applied to a different number of filters; as mentioned before, Pols *et al.* had used 18 filters, Klatt, 36 filters, and Assmann and Summerfield used as many as 75- or 125-channel frequency representations. However, in this study, 32- and 64-channels for frequency analyses are used. 128-channel analysis is thought to be excessive. Thus, for Euclidean metric, 32-channel frequency analyses are carried out. This is compatible with the number of frequency channels used in previous acoustic distance analyses. For WSM and WN2DM, 64-channels are used.

It must also be noted that, although the degree of correlation improved for the

optimised WSM in Klatt (1982a), the improvements were modest in Assmann & Summerfield (1989). Therefore, to start with, the constants for the weighted distance metrics are set to unoptimised values: $k_E = 0$, $k_{Gmax} = k_{Lmax} = 10e^6$, and metrics are referred to as **Slope** and **N2D** metrics.

### 3.2.4 Auditory space

In order to obtain the auditory space of the stimuli, the acoustic distance matrices obtained from the different distance metric analyses were used as input to the nonmetric MDS analyses. The output of the MDS analyses was the set of coordinate points of each auditory dimension. The data were also analysed by principal components analysis (SAS version 6), for Euclidean metric. Since the plots of both processes were almost identical, and since the MDS process could be used to obtain acoustic maps for all the other metrics, only the MDS results are described here.

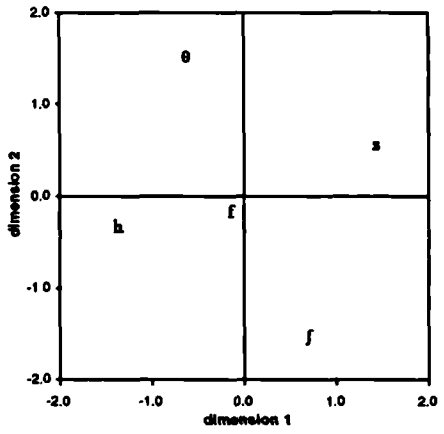
1- and 2-dimensional MDS solutions were tried out; 3-dimensional analyses of the data were not always permitted by Proc ALSCAL. The values for badness-of-fit were around 0.2 for the 1-dimensional solutions, but the 1-dimensional plots were not interpretable. 2-dimensional solutions provided almost perfect fit (badness-of-fit was almost 0) for the distance matrices.

The resulting auditory configurations are presented in Figures III-18 to III-19.

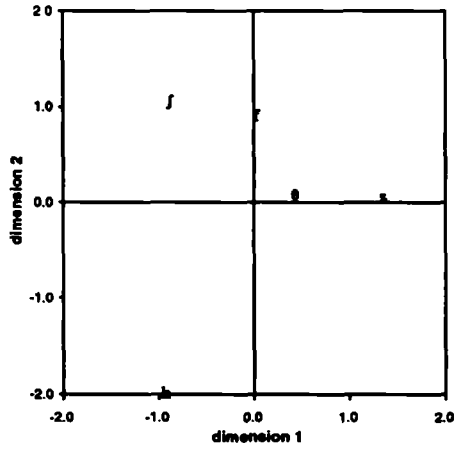
Figures III-18 (a) to (c) show the auditory dimensions of the cut-out fricative segments used in set 1, analysed by the three different distance metrics. The Euclidean space shown in Figure III-18 (a) seems to be strongly related to the perceptual map in Figure III-13, especially for dimension 1. For dimension 2 of the acoustic space, the positions of [f] and [h] do not correspond well to the perceptual configurations.

The acoustic configurations obtained by the other two distance metrics do not correspond well to the perceptual organisations (Figures III-18 (b) and (c)).

(a)




(b)




(c)

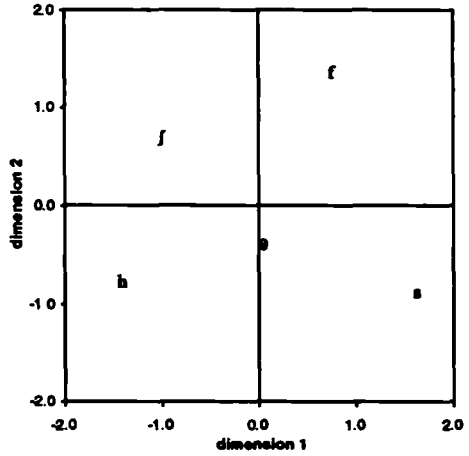



**Figure III-18.** Acoustic maps for the cut-out natural fricative segments based on three different distance metric analyses. (a) Euclidean (b) Slope (c) N2D.

Figures III-19 (a) to (c) are obtained from the MDS analyses of the distance matrices based on the different distance metric analyses, for the LPC synthesised fricatives. Comparing these with the perceptual configurations in Figure III-14, it can be observed that dimension 1 of the Figures III-19 (b) and (c) bears a strong resemblance to that of the perceptual map. However, the most striking overall feature is that the configurations obtained from the N2D look extremely similar to the perceptual plot.

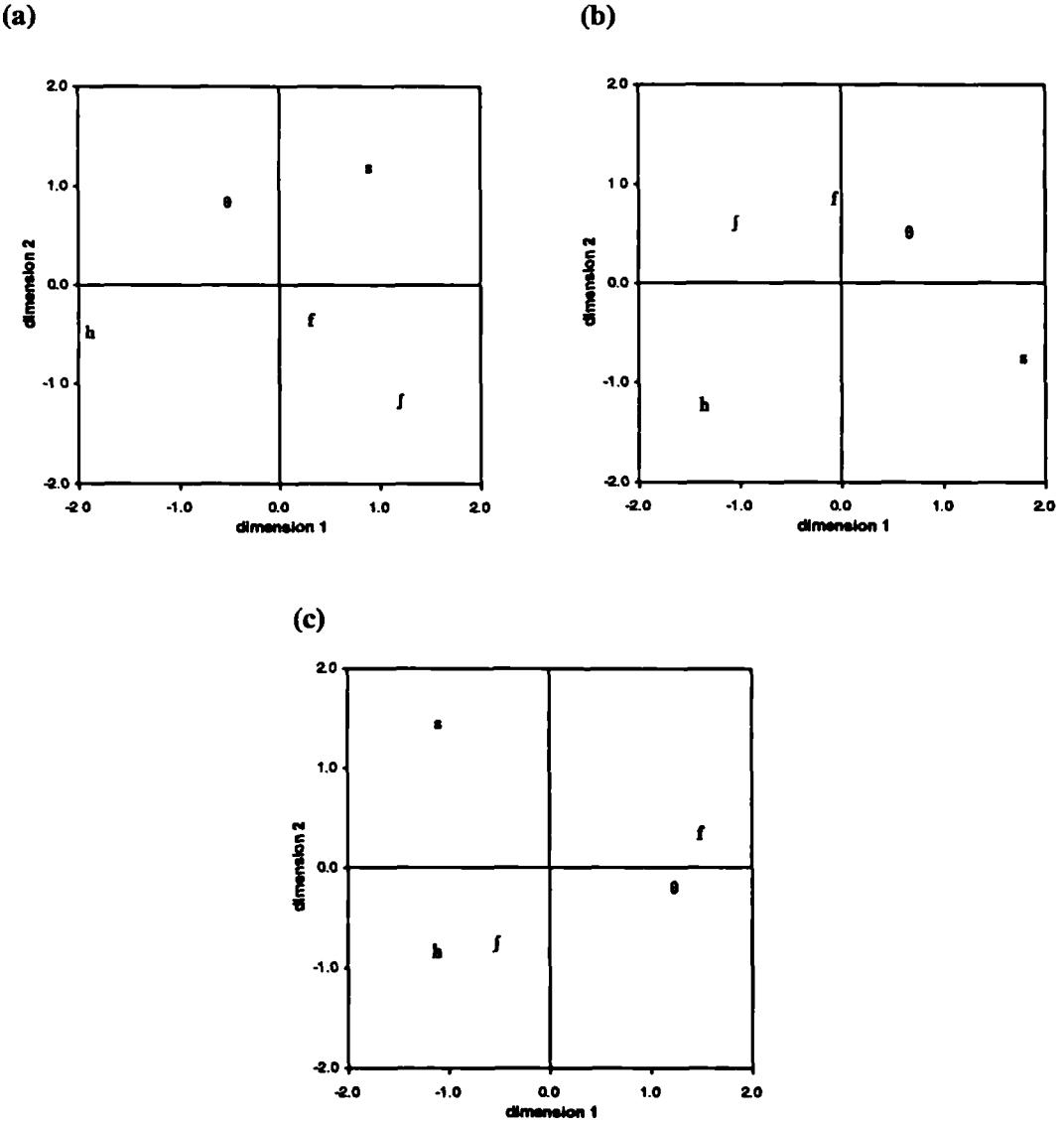



**Figure III-19.** Acoustic maps for the LPC synthesised fricatives based on three different distance metric analyses. (a) Euclidean (b) Slope (c) N2D.

Now auditory spaces have also been established, and we have made some informal comparisons with perceptual spaces. In the next section, the perceptual and auditory spaces will be compared quantitatively, by canonical correlation analysis.

### 3.3 Comparison between the perceptual and auditory spaces

Describing the relationship between the perceptual and auditory dimensions as above provides a useful overview. A technique which allows a more quantitative evaluation of the relationship between two spatial presentations is **canonical correlation analysis.** Canonical correlation is a statistical technique for analysing the association between two sets of variables. This is done by forming linear combinations of the variables that have maximum correlation. How this is done is best illustrated by a simple example. Suppose we have two variable sets (the 'dependent' and 'predictor' sets), both arranged in a pencil shape, as below in Figures III-20 (a) and (b).

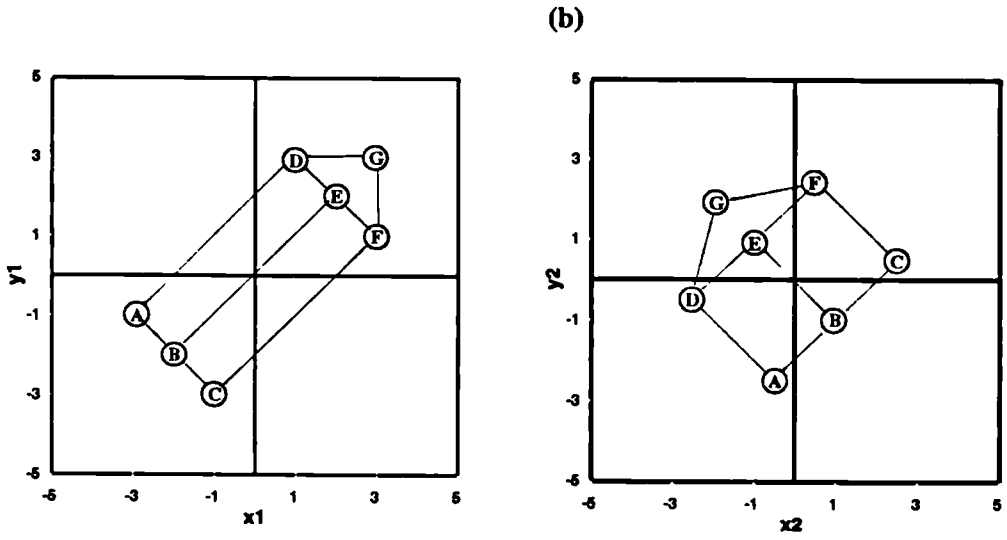



**Figure III-20.** Two hypothetical variable sets, (a) dependent set, (b) predictor set.

Canonical correlation analysis between these two sets of variables by SPSS (windows version 6.1) produces canonical coefficients of 1.000 for dimension x and 0.992 for dimension y. The analysis also tests zero probability for given correlations, and this indicates their **significances.** For this particular example, the correlations were highly significant of $p<0.001$. **Canonical scores** are the new sets of coordinates which were scaled

and rotated to give optimal correlations. These are used to give a graphic representation of the relationship between the variables measured. These are tabulated below:

| | x1 | y1 | x2 | y2 | new x1 | new y1 | new x2 | new y2 |
|---|---|---|---|---|---|---|---|---|
| A | -2.896 | -0.996 | -0.469 | -2.499 | 1.06 | 1.01 | 1.11 | 1.00 |
| B | -1.997 | -1.993 | 0.987 | -0.989 | -0.12 | 0.87 | -0.11 | 0.80 |
| C | -0.992 | -2.985 | 2.493 | 0.459 | -1.36 | 0.69 | -1.32 | 0.64 |
| D | 0.997 | 2.896 | -2.492 | -0.499 | 1.30 | -0.68 | 1.33 | -0.62 |
| E | 1.994 | 1.991 | -0.978 | 0.899 | 0.12 | -0.87 | 0.13 | -0.75 |
| F | 2.998 | 0.984 | 0.488 | 2.411 | -1.13 | -1.04 | -1.09 | -0.95 |
| G | 2.978 | 2.978 | -1.919 | 1.899 | 0.18 | -1.30 | 0.21 | -1.54 |

**Table III-6.** Coordinates for the dependent (x1 y1) and predictor (x2 y2) sets, with canonical variables (new x1 y1 x2 y2).

The canonical scores are plotted in Figure III-20 (c).

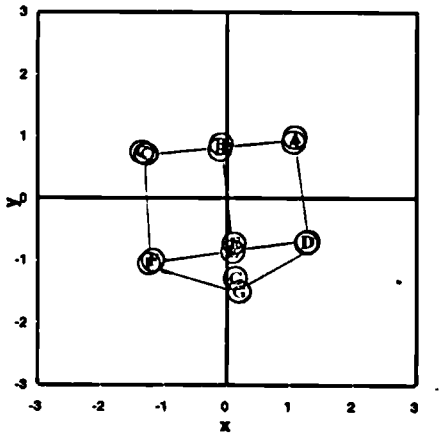



**Figure III-20** (c). A plot of canonical scores (new set of coordinates) showing perfect match between the two sets of variables shown in Figures III-20 (a) and (b).

Since the variables under comparison have the same basic shape and their centres of gravity coincide in the origin, the correlation is almost perfect. In a real situation, the predictor sets may contain more variables than the dependent set. In that case, canonical correlation finds variables in the predictor set which result in maximum correlation with the variables in the dependent set. Therefore, this method is appropriate for finding

correlations between two different coordinate sets in multidimensional space. One theoretical limitation is that an orthogonal relationship is assumed between the compared variable pairs. Thus, when the compared variables are not completely independent of each other, it should be borne in mind that the significance of the correlations may be a little less than indicated (Schiffman *et al*, 1981: p191). Apart from this, the analysis is the most appropriate for investigating the associations between corresponding coordinate points in two multidimensional spaces.

The coordinates of perceptual and auditory spaces were compared by applying canonical correlation analysis, and the results are presented in Tables III-7 and III-8. However, canonical scores are not considered at this stage, since the data are not large enough to carry any statistical significance.

| Distance metrics | Statistics | Dimension 1 | Dimension 2 |
|---|---|---|---|
| Euclidean | Canonical correlation | .984 | .641 |
| | Probability | .254 | .359 |
| Slope | Canonical correlation | .994 | .044 |
| | Probability | .205 | .956 |
| N2D | Canonical correlation | .830 | .487 |
| | Probability | .737 | .513 |

**Table III-7.** The canonical correlation values and the probability levels, for each of the dimensions between the perceptual and physical data of cut-out fricatives.

As predicted, the probability levels for the null hypothesis — that all the canonical correlations are 0 in the population — are not significant; this means that the sample size is not large enough to draw any definite conclusions. However, the canonical correlation may be strong enough to support the qualitative observations made previously. This was the case for the Euclidean metric analyses. Also, the fact that dimension 1 was well correlated for Slope metric but not dimension 2, was well illustrated by the canonical correlation coefficients.

| Distance metrics | Statistics | Dimension 1 | Dimension 2 |
|---|---|---|---|
| Euclidean | Canonical correlation | .626 | .299 |
| | Probability | .935 | .701 |
| Slope | Canonical correlation | .954 | .692 |
| | Probability | .385 | .308 |
| N2D | Canonical correlation | .996 | .816 |
| | Probability | .104 | .184 |

**Table III-8.** Canonical correlations between the perceptual and acoustic spaces for the LPC synthesised fricatives, based on the three different distance metrics.

The canonical correlations shown in Table III-8 also confirm the qualitative observations made earlier, that N2D acoustic space can be matched to the perceptual configurations rather well.

Note that, for the stimulus set 2, we achieved progressively better predictions of the perceptual data as we move from the simple Euclidean metric (which gives equal emphasis to the spectral peaks as well as to the valleys) to the more sophisticated metrics (which tend to emphasise spectral peaks and shoulders in the spectra, as shown in previous studies) (refer back to Table III-5). However, this was not the case for set 1. Instead, Euclidean metric gives highest correlations, although they are not very high. Since we did not have a large enough data set to reach the required significance level, the results were taken only as an indication of the relationship. A more detailed comparison between these metrics will be carried out in the main auditory distance modelling experiments (Chapter V).

## 3.4 Discussion

Purpose of the work in this chapter was to make preliminary investigations of the spatial relationship between phonetic, perceptual and auditory domains, for fricative sounds. As in the vowel studies (§II.2.1), the perceptual dimensions of the cut-out fricatives (stimulus set 1) were readily related to the known phonetic dimensions 'place' and 'sibilance', but for the LPC synthesised fricatives (stimulus set 2), the relationship between the perceptual and phonetic dimensions was not so obvious.

The canonical correlation analysis showed that both perceptual dimensions 1 and 2 of LPC synthesised fricatives correlated closely with their auditory dimensions (0.996, 0.816). However, for the cut-out fricatives, the correlation coefficients for the first two dimensions between the perceptual and auditory spaces were 0.984 and 0.641, which means that dimension 2 was not well correlated. This contradicts the result obtained by Klein *et al.* (1970), with the natural 100 ms segments of Dutch vowels. Their correlation values reported were 0.997, 0.995, and 0.907, which indicate an excellent matching between the first three dimensions of the perceptual and physical spaces.

Now, our understanding so far of the general relationship between the perceptual and auditory spaces can be restated as follows:

- For synthesised stimuli, based on simplified steady-state spectral shape, there has been very close matching between perceptual and auditory spaces, in both vowels and fricatives.

- For the natural segments, the perception of fricatives could not be fully explained by their auditory spaces, and this is different from the results obtained from the vowel stimuli of a comparable nature.

The relationship between the phonetic and the perceptual spaces could be summarised as follows:

- For the synthesised stimuli, phonetic interpretations were not so clear-cut, in both vowel and fricative data.

- For the natural segments, the most prominent feature of the perceptual map was that the fricatives were clearly distinguishable according to their place of articulation feature, and vowels could be distinguished according to front/back and tongue height features.

The fact that the 'place' feature was clearly observable in the perceptual maps of

the natural fricative stimuli of experiments 1 and 2, but not in the synthesised stimuli, could mean that the synthesised fricatives may not have been perceived as fricatives, but as meaningless nonspeech sounds. This means that the mismatch between the perceptual and auditory spaces for the natural fricative segments may be indicative of something fairly exclusive about speech perception. That is, the relationship between the perceptual, auditory and phonetic domains of the natural fricatives seems to support arguments for the existence of speech perceptual mechanisms that incorporate phonetic/phonological knowledge, apart from the acoustic information contained in the spectra. If this is correct, it is necessary to investigate the extent to which the linguistic knowledge of a hearer would distort the speech signal, compared to the perception of nonspeech sounds. This issue will form the theme of subsequent investigations in chapters IV to VI.

Before we can suggest strategies for addressing this issue (in §5), we need to verify the relationship between perceptual and auditory spaces for nonspeech sounds. If we postulate that there might be some kind of phonetic interference in the perception of speech sounds, this hypothesis assumes that the perceptual pattern of nonspeech sounds of comparable quality is completely explicable by the auditory spectral analyses. Therefore, the next section is devoted to the verification of this assumption.

## 4 Experiment 3: Shaped white noises

### 4.0 Introduction

The above assumption is tantamount to saying that the canonical correlation of perceptual and auditory spaces of simple noises should be very high. To test this hypothesis, the stimuli were drawn from two-formant filtering of white noise so that the acoustic structure of the sounds can be plotted in a two-dimensional space. If the hypothesis is confirmed, the perceptual and auditory spaces of the sounds should also be in two dimensions, in which the noises are positioned in a similar fashion to the resonant frequencies.

### 4.1 Stimuli design

Since there are five fricatives in English, the number of stimuli was set to five. Each stimulus was white noise filtered through two resonators with fixed bandwidth of 1000 Hz but with various centre frequencies. To determine the centre frequency of each

resonator a two-dimensional grid was constructed, as shown in Figure III-21, so that no pair of resonator peaks could occur at the same frequency and the noise c is equidistant from the others. In effect, each noise has two 'formants' F1 and F2.

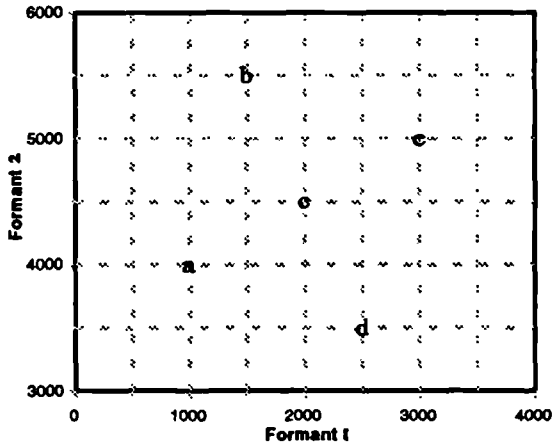



**Figure III-21.** Two dimensional grid of F1 against F2 (the design of stimuli).

The auditory intensity of each stimulus at the loudspeaker was normalised using a speech pressure level meter. The duration of each stimulus was 0.5 seconds.

## 4.2 Subjects

Ten students, studying B.Sc. in Speech Science, volunteered to listen to the stimuli. All of them were native speakers of English, and had normal hearing. Their perceptual task was the same as the previous experiments (triadic comparison).

## 4.3 Analyses

Procedures for perceptual and auditory analyses were the same as for experiment 2. Perceptual space was obtained from nonmetric analysis. The two-dimensional space was chosen as the solution. Badness-of-fit for the 1-dimensional solution was 0.28 and for the 2-dimensional solution, it was 0.12. 3-dimensional analysis was not permitted.

Given the substantial increase in the number of subjects participating in this experiment, compared to the previous two experiments, it would be sensible to discuss the subject weights. This is shown in Figure III-22. The mean of the subject weights for

dimension 1 was 1.02, and for dimension 2 it was 0.96, with only one of the ten subjects showing an outlying value (subject 2). This subject placed more weight on dimension 1, but still the two-dimensional solution was adequate for this subject. The similar mean weights for the two dimensions mean that the subjects had placed similar emphasis on the two perceptual dimensions.

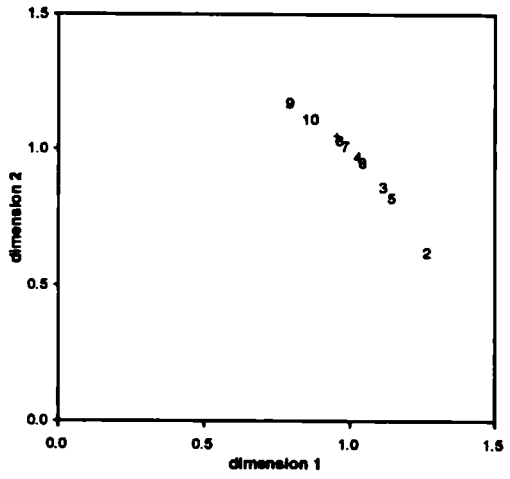



**Figure III-22.** Subject weights for dimensions 1 and 2.

Three different auditory spaces, Euclidean, Slope, and N2D, were obtained as described in §3.2, and the numerical and graphic correlations of each auditory space with the perceptual space is examined in the next section.

## 4.4 Canonical correlations between perceptual and auditory spaces

Perceptual and auditory coordinates were used as inputs to canonical correlation analysis. The canonical coefficients and their significance are shown in Table III-9, for the three different distance metrics.

| Distance metrics | Statistics | Dimension 1 | Dimension 2 |
|---|---|---|---|
| Euclidean | Canonical correlation | .972 | .934 |
| | Probability | .115 | .079 |
| Slope | Canonical correlation | .994 | .794 |
| | Probability | .087 | .222 |
| N2D | Canonical correlation | .987 | .710 |
| | Probability | .160 | .305 |

**Table III-9.** Canonical correlations between the perceptual and Euclidean, Slope and N2D auditory spaces of noises.

The correlations show that Euclidean space can adequately predict the perceptual organisations of the noises. For the other metrics, the correlations are high for dimension 1, but not for dimension 2. A graphic display of correlations may clarify the results.

Figures III-23 (a) to (c) are the plots of the new sets of coordinates (canonical scores), which were scaled and rotated to give optimal correlations between the perceptual and auditory configurations. It is absolutely clear from Figure III-23 (a) that the two configurations are very closely related for the Euclidean metric. For the other metrics, the matching between the two coordinates is obviously not so good. This is in accordance with Klatt's claim (1982a) (§II.2.3), in that, for 'psychoacoustic judgments' (as opposed to the 'phonetic judgments'), other spectral information besides the formants centre frequencies, like spectral tilt and amplitude properties, may also be needed to account for the perceptual judgments. On the other hand, since the noises have prominent formants, Slope and N2D metrics might have performed better than the Euclidean metric. A possible explanation is that the Slope and N2D metrics, which overly concentrate on peaks, were better predictors of perceptual distances between two very similar spectra (as in Klatt, 1982a; Assmann & Summerfield, 1989; Kewley-Port & Atal, 1989), but for these noise spectra, which are very different, the metrics may not give accurate distances between the spectra. However, this is by no means conclusive evidence for the performance of these distance metrics, and the metrics are investigated further in Chapter V.

The main point to note in this experiment is that for noises, the perceptual and

Euclidean auditory spaces were highly correlated. If we reiterate this result for nonspeech sounds in general, the perception is a simple one-to-one process between the acoustic signals and the percepts.

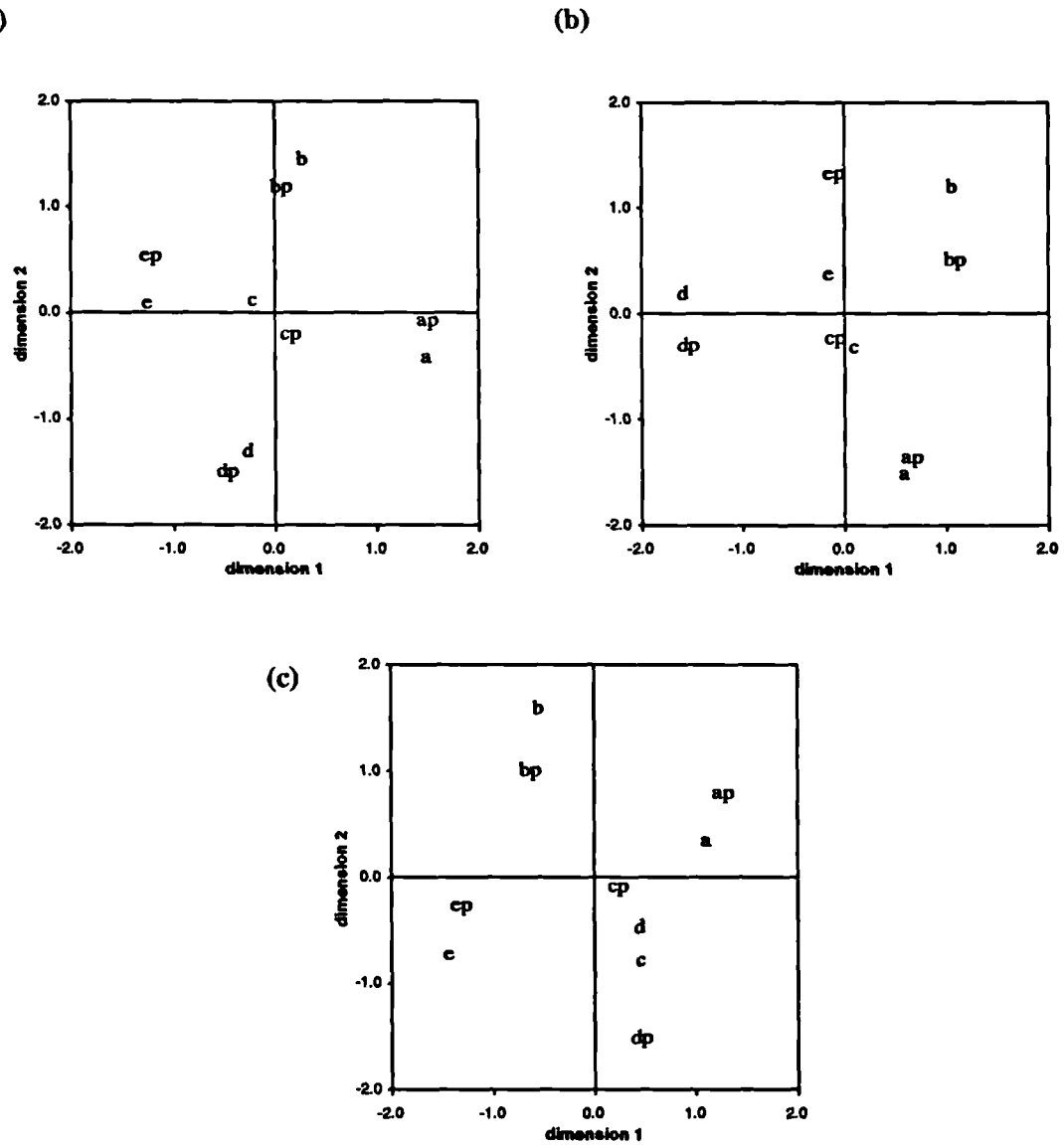



**Figure III-23.** Plots of canonical scores for perceptual and (a) Euclidean (b) slope (c) N2D spaces. (Noises with p refers to the perceptual coordinates while ones without p refers to auditory coordinates).

## 5 Pointers for future experiment design

Now we have shown that the perception of nonspeech sounds is fully explainable by the auditory modelling of these sounds, we can go back to the issue that was raised at the end of experiment 2 (in §3.4) — whether the fricatives were perceived differently from the

vowels, incorporating higher level linguistic knowledge, other than the acoustic information in spectra. If so, this needs to be further investigated.

The main strategy we will adopt is that the relationship between phonetic, perceptual and auditory spaces is examined not only within a set of stimuli, but for an entire range of fricative sounds, from natural syllables down to simple two-formant sounds. This means that we can examine the spatial relationships *vertically* as well as *horizontally*. As we go down the stimulus list vertically, from complex natural stimuli to simple two-formant stimuli, changes between horizontal acoustic and perceptual configurations will be monitored by MDS analysis, so that it will be possible to observe whether:-

- The perceptual mapping will change gradually from stage to stage, corresponding to the gradual simplification of the acoustic properties of the fricatives or;

- There may be an abrupt change in perceptual mapping, despite of the gradual changes in acoustic characteristics of the materials, reflecting perceptual 'switch' from speech to nonspeech mode.

These experimental observations are possible on the assumption that MDS technique is capable of reflecting such perceptual and auditory changes. The acoustic complexity of natural fricative syllables will be gradually simplified by taking into consideration the acoustic cues, which were mentioned in §II.4, such as vowels, and transitions, as well as simplifications of fricative spectra themselves. This means that auditory distance metrics also need to cater for 'dynamic' spectral changes, and this method may eventually be extended to auditory modelling of stop consonant perception elsewhere. The statistical validity of both perceptual and acoustic data needs to be confirmed since the group of subjects involved in experiments 1 and 2 numbered only five. Finally, the auditory analyses were based on the productions of one speaker only. Speaker- or subject-specific factors are investigated in multiple speaker production tests (Chapter VI).

# *Chapter IV. Perception tests*

## 1 Introduction and objectives

In Chapter III, we explored the matching between perceptual, phonetic, and auditory organisations of two sets of fricatives; one comprised the cut-out natural fricatives, the other consisted of synthetic fricatives. The matching between auditory and perceptual domains was much higher for the synthetic materials than the natural fricatives. On the other hand, the perceptual organisations of natural fricatives could be given clear phonetic/linguistic interpretations, which was not the case for the synthetic fricatives. This raised questions as to whether different processes might have been involved in speech versus nonspeech perception (§III.5). If the auditory analysis (in particular, the N2D metric) adequately modelled the perceptual processes of the synthetic stimuli, then the mismatch between the perceptual and psychoacoustic planes of the natural speech materials could represent a phonetic/phonological influence peculiar to speech perception.

In order to investigate this potential phonetic influence, we need to investigate the relationship between the auditory and perceptual matching of the intermediate sounds, between the synthetic materials and natural speech. At what point does the auditory distance modelling become inadequate for representing the perception of these sounds? Or would the mismatch between the two domains gradually increase from nonspeech to speech? Would it make any difference whether the subjects heard the same stimuli in speech mode or nonspeech mode?

This chapter discusses the designs of stimulus sets and listening procedures appropriate to the above questions, and reports the results of perceptual tests. Phonetic interpretations of the perceptual dimensions arising from an MDS analysis of these results are attempted. Subsequent auditory modelling and assessment of suitable distance metrics of the stimuli are described in the next chapter.

## 2 Design of stimuli

In the preliminary analyses (Chapter III), we have tested four different sets of materials:

1. Whole syllables: /fɑ θɑ sɑ ʃɑ xɑ çɑ hɑ/
2. Cut-out fricatives: /f θ s ʃ h/
3. LPC synthesised fricatives with 10 coefficients: /f θ s ʃ h/
4. Two-formant white noises: a b c d e.

Following on from the results of the perceptual tests in the pilot experiments and previous studies on fricative perception, seven stimulus sets were constructed so that they reflect a change in terms of their acoustic parameters from natural speech to evidently synthetic versions. Appropriate changes have been made to the four sets used in the pilot experiments and additional stimulus sets were also introduced. Details of stimuli and the rationale for each stimulus set design are given below.

The initial stimuli were fricative syllables /fɑ θɑ sɑ ʃɑ xɑ hɑ/, read by a female English phonetician, native speaker of R.P., in a falling tone. [xɑ] was added to the usual English fricative set to increase the number of combinations in similarity judgments[1]. The speaker was seated in an anechoic chamber. A Bruel & Kjaer sound level meter type 2231 and micro phone (Bruel & Kjaer, type 4165) were used for recording the speech pressure signal, placed half a metre from the mouth and 15 degrees from the median plane. The materials were recorded onto a Sony DTC-1000ES digital audio tape recorder. They were digitized with a 20 kHz sampling frequency and 16-bit quantization rate, and transferred onto computer disk. From these initial stimuli, henceforth referred to as **whole syllable** stimuli, six additional stimulus sets were devised. Despite careful recording conditions, these whole syllable stimuli had to be normalised with respect to their loudness, since some of the stimuli were evidently louder than the others, thus providing an extraneous cue to a particular stimulus. Loudness of the stimuli was measured again by an artificial ear (Bruel & Kjaer, type 1453, microphone type 4134), but it was difficult to obtain consistent readings. The main cause for this was attributed to the difficulty in

[1]Since /x/ occurs in the Scottish word 'loch', it is not completely foreign.

fitting the ear piece (of a headphone) to the artificial ear; slightest movement caused fluctuations in the meter reading. In the end, loudness normalisation was done on the basis of RMS values of 200 ms of the loudest part of the vowel. As it was mentioned in §II.4.1, relative amplitude of frication in relation to the vowels may be an important cue for fricative identity, and this was therefore maintained. The resulting stimuli (just the vowel parts) were checked and it was found that they had no discernable differences in loudness.

From these whole syllables, transition parts were removed in the next set of stimuli which were called **no transition** stimuli. This was done in order to observe the perceptual effect of transitions, as the evidence for the importance of transitions proved contradictory (§II.4.3). The transition was taken out in two stages; first by cutting the fricative part out from the whole syllable stimuli, and then, by gluing on a synthetic vowel section straight after the fricative section. The fricative part was cut by an interactive computer program (SFS, written by Mark Huckvale, University College London). The beginning of transition was identified by both wide- and narrow-band pass filtering. In order to ensure that exactly the same acoustic parameters were involved in the vowel section for each stimulus, an /ɑ/ vowel was synthesised using a Klatt formant synthesizer (KPE, written by Andrew Simpson, UCL). The parameter specifications were based on the measurements of natural tokens produced by the same speaker. The amplitude level of the synthetic vowel was adjusted to a level comparable to that of each corresponding natural vowel. Also the onset of the vowel was adjusted to give a smooth build up from the fricative segments, (but without formant transition).

Since the most important perceptual cues for fricatives are thought to be contained within the fricative spectra, the next stimulus set consisted of the fricative sections only, cut out from the whole syllables; hence, these are called **cut-out** stimuli. Although it is already shown in §III.3 that the fricative portions contained sufficient information for correct phonetic discrimination of the stimuli, statistical significance was not obtained. Also, further acoustic parameters were simplified. These were intensity and duration of the fricatives. In the previous two stimulus sets, loudness was controlled with respect to the vowels, but not the fricatives. Intensity was normalised by adjusting the RMS levels of fricative to an equal level. The duration was normalised to 200 ms, with the help of a computer program ('respeed', written by Mark Huckvale), which can speed-up or slow-

down a piece of speech without changing the pitch[2]. This was because the length of the fricatives may have some effect on perceived loudness values for fricatives. As reviewed in §II.4, fricatives may be perceived differently depending on their amplitude levels.

In the next stimulus set, the dynamic spectral properties were taken out, so that the stimuli have the same cross-sectional spectral shape throughout the whole length. This was done by LPC synthesis, modelled on the spectral cross-sections of the natural fricatives. We saw in §III.3 that the perceptual map of LPC synthesised fricatives with 10 coefficients could not be related to any recognised phonetic features; in response, the coefficient number was doubled to model the fricative spectral characteristics with greater accuracy. The number of coefficients was set to 22 — this number was usually used to model vowel spectra by LPC, and most of the spectral characteristics which can be modelled by spectral peaks and valleys should be reflected in the spectral shape. The desired effect is like duplicating a period out of vowels many times to obtain static spectra in Pols *et al.* (1969). The length of stimuli were always set to 400 ms. These synthetic fricatives are also normalised with respect to overall RMS level The stimuli in this set were called **LPC22** stimuli.

Since the LPC synthesised fricatives in §III.3 might have been perceived as nonspeech sounds, the synthetic vowel used in the 'no transition' set was attached to synthetic fricatives with LPC 10 coefficients, in order to induce *speech mode* perception. The stimuli might otherwise be treated as noises rather than synthetic speech. The spectral shape of the fricatives is further simplified; this time 10 coefficients were used. The duration of the fricative portions was set to 400 ms and loudness was normalised by RMS level adjustment. These stimuli were called **LPC10a.**

The effect of a synthetic vowel in perception was compared to stimuli without the vowel in the next set. Other acoustic parameters remained unaltered. These stimuli were called **LPC10** stimuli.

In the final stimulus set, fricative spectra are modelled by LPC with just four coefficients, so that the fricative spectrum is characterised by two peaks. The perception of these stimuli can be directly compared to the results of the experiment in §III.4, where

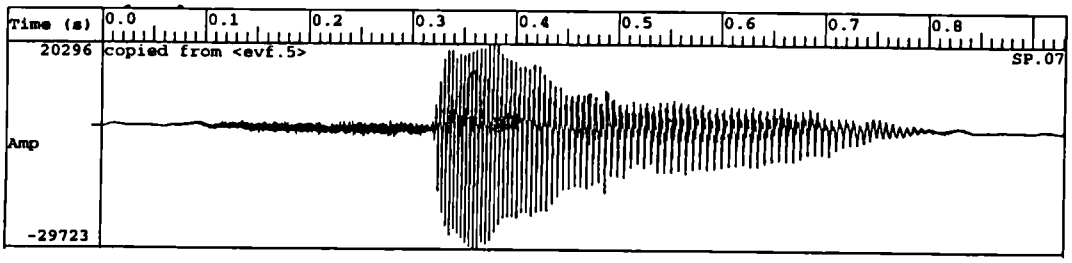
[2]Exactly how it works is beyond the scope of this study, but for those who are interested, it was based on the SOLA algorithm of Roucos & Wilgus (1985).

perceptual and auditory spaces of two-formant white noises were compared. Would there be any differences between the perception of two-formant spectra, one modelled on speech, the other on just any two formants? These stimuli were called **LPC4** stimuli.

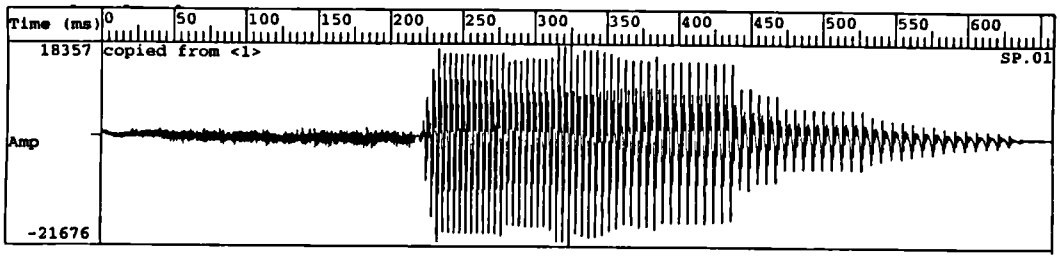
A summary of stimulus sets is given in the box below:

1. Whole syllable: natural productions of /fɑ θɑ sɑ ʃɑ xɑ hɑ/
2. No-transition: fricative section + synthetic /ɑ/
3. Cut-out: fricative portions with intensity and duration normalisation
4. LPC22: fricatives synthesised with 22 coefficients
5. LPC10a: fricatives synthesised with 10 coefficients with synthetic /ɑ/
6. LPC10: fricatives synthesised with 10 coefficients
7. LPC4: fricatives synthesised with 4 coefficients

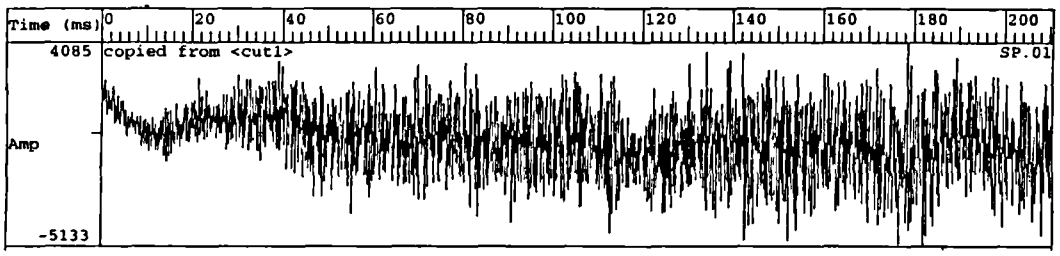
For each stimulus set, an example of the speech signal is given in Figures IV-1 (a) to (g). Spectrograms and spectra of the stimuli are shown in the Addendum.



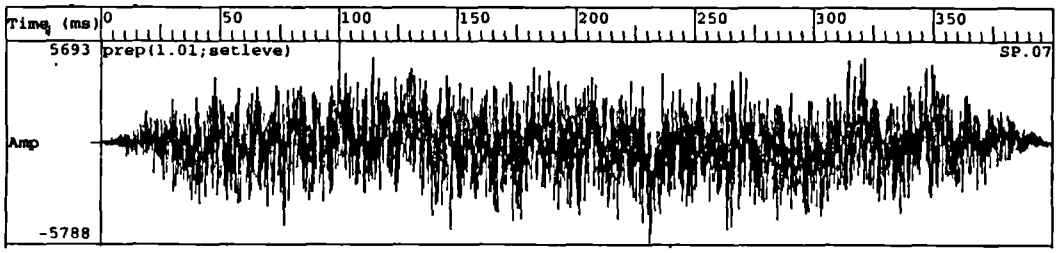

(a) /fɑ/ in whole syllable set.



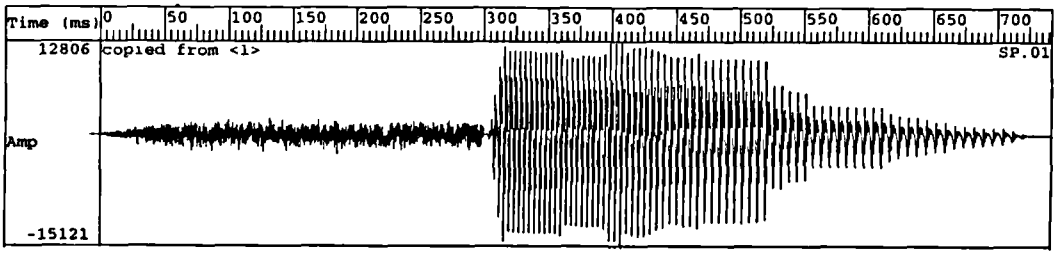

(b) /f/ followed by a synthetic vowel, /ɑ/ in no-transition set.



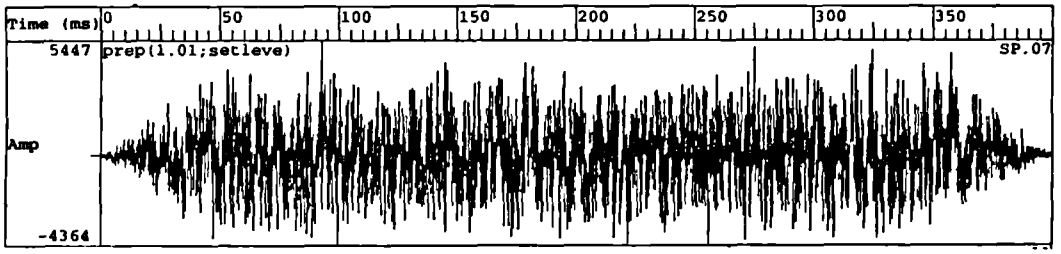

(c) /f/ segment in cut-out set.



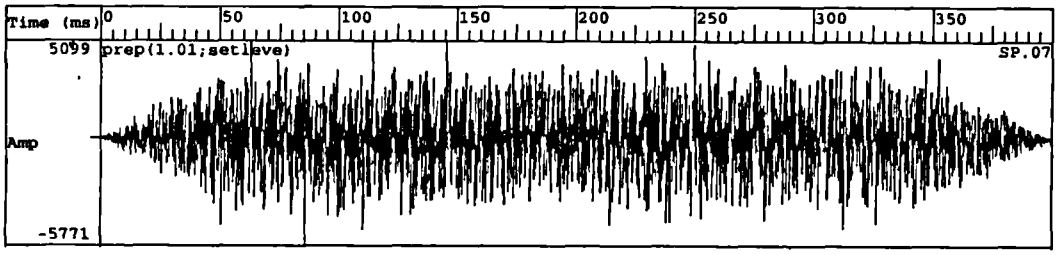

(d) /f/ in LPC22 set.




(e) /f/ with following /a/ vowel in LPC10a set.




(f) /f/ in LPC10 set.




(g) /f/ in LPC4 set.

**Figure IV-1.** Examples of stimuli used in the perception tests.

These stimuli were recorded back to digital tapes in a form appropriate for testing. How this is done is described in the following section.

## 3 Subjects and procedure

First, each stimulus in each stimulus set was paired with two other stimuli in the set, giving 60 (= 6×5×4/2) possible stimulus triads. The first stimulus was repeated after the second stimulus so that the listeners were asked to choose whether the members of the first pair or the second pair sounded closer to each other. Two sets of stimuli were presented in each session. Half of the stimuli were presented in the order AB AC while the other half was presented in the order AC AB. The sequence of triads was randomised for each presentation to take into account the ordering effect.

20 students, studying B.Sc. in Speech Science, were paid to listen to the stimuli. All were native speakers of English, and reported normal speech and hearing.

Subjects were split into four groups. The first two groups (henceforth, called subject **Group A**) heard the stimuli sets in the order of whole syllable set to LPC4 set. The other two groups (subject **Group B**) heard the stimuli sets in reverse order, that is, from LPC4 set to whole syllable set. This order was used to

i) cancel out any learning effect, and

ii) address the issue that different modes may be used for processing and interpreting different types of stimulus. In this case, previous exposure to synthetic speech might affect the triggering of the speech mode of perception for Group B (Repp, 1982).

Subjects were tested five at a time in a sound treated room. Testing lasted for 1/2 an hour on each of four days, over a period of four weeks. Day 1 began with practice materials: they consisted of the six test stimuli sounds in the order of /f θ s ʃ x h/, and the material for the first main test, but in different random orders. (These were trials, and they were not taken into account in the analysis of the results.) The main experimental task consisted of listening to two sets of stimuli. Each test set began with a list of the stimuli set and 5 trial pairs to start them off. There was 0.1 second of inter-stimulus and inter-stimulus-pair gap and 2 seconds pause after two pairs were presented. There was a pause of ten seconds after each block of 5 similarity judgement pairs. After that pause the listeners were prompted by a tone for the next block. Each subject made judgments on a

total of 420 triads over all sessions.

Subjects were given both written and verbal instructions as to how to respond to the stimuli pairs, without mentioning any criteria for making the similarity judgements, as described in §III.2.3.

Instructions described the fricatives only as English fricatives /f θ s ʃ h/ and the fricative occurring in the Scottish word 'loch', for the voiceless velar fricative /x/. The members of subject Group A were comfortable with these instructions at the outset, but complained that some of the stimuli sounded very similar to each other from the LPC22 set and consequently, that it was difficult to make choices. They were reassured that there are no right or wrong answers, and were asked to choose whichever pair sounded more similar, without worrying about the original linguistic identity of the segments. The subjects in Group B complained straight away, saying that the stimuli did not sound like English fricatives at all. They were assured that the stimuli were meant to sound very artificial and told to base their judgments on 'similarity' of stimuli, nothing else. Although subjects complained at various stages of the experiments, none of the subjects trailed behind and all of the subjects completed all four sessions.

Similarity data were accumulated in the same way as before: the pairs selected to be more similar were assigned 1 scores, and the pairs which were not selected, 0 scores. In this way, a matrix of data indexing the perceived relationships among the six stimuli was obtained for each subject (listed in Appendix).

ALSCAL, because of its versatility of program options, which allows both metric and nonmetric, 2-way and 3-way multidimensional scalings, was used to analyse the perceptual data. It was previously reported that, when the metric option of ALSCAL is used, the subject and stimulus spaces give configurations similar to those given by the INDSCAL analyses. However, at the nonmetric level — that is, when the data are viewed as being at the ordinal level as supposed to the interval level in the metric option — ALSCAL was reported to have a tendency to compress the distances among stimuli (as mentioned in §III.2.6). To check the stability of the configurations and interpretability of the different solutions, both metric and nonmetric 2-way and 3-way ALSCAL analyses are carried out.

The proc-ALSCAL (SAS system) allows the input to be asymmetric similarity

matrices[3]. Thus, the 'square' matrix input was used so that the entire matrix, without separate symmetrisation as we have done in pilot experiments, could be analysed.

## 4 Results and discussion

It has not been customary in MDS studies of speech sounds to explicitly study the subjects' ability to behave in a consistent way — especially so, if the third mode of 3-way MDS comprises of individual listener's data. We have already seen in §4 how individual differences are accommodated within WMDS. The conformity of each subject's data to the data of group as a whole is indicated by the weighting of each subject on each dimension. Since each subject data is treated independently of the others in WMDS, noise caused by subject inconsistency in the data will cause difficulties in convergence, and the derived spatial solutions will be unstable.

In order to ensure the stability of any given MDS solution, a comparison of spatial solutions given by two or more different types of MDS analysis is carried out. Here it is mainly the results of metric and nonmetric ALSCAL analyses that are compared. Although the results are not presented, solutions of 2-way, unweighted, MDS were also examined for this purpose. As an indication of the similarity between two spatial representations, canonical correlation analyses were carried out where these were considered appropriate.

Another customary method of measuring reliability of data in MDS analysis is to compare stimulus spaces from two split-halves of subject data. If the MDS solution of one-half of the sample is similar to that of the other half, the data as whole is thought to be reliable (Fox *et al.*, 1995). The results of Group A were compared with the corresponding data of B. Since the subject Groups A and B had heard stimuli set in the opposite order — A, from the whole syllable set to LPC4, but, B, vice versa — different listening modes could have influenced the results of two groups. To safeguard against this, when a mismatch occurs, the results from Groups A and B were interpreted separately and a further split-half analysis was applied.

The data matrices for each listener for each stimulus set are presented in Appendix. The subjects are labelled from 1 to 20. An initial look at the content of the matrices shows that the data are fairly symmetrical, which contributes to the reliability of individual

[3]SPSS allows input matrices to be asymmetric only if they are distance scores.

subject judgments, although substantial asymmetry is not uncommon in similarity judgment data. Also, an informal examination of the entries in the matrices shows that they conform to intuitive expectations based on phonetic descriptions of fricatives; that is, for most of the subjects, /f/ is judged more similar to /θ/ than it is to /h/. Further inspection of the raw data would reveal additional suggestive patterns; and detailed statistical analyses may also show some indication of how a single subject varied across stimuli sets. But this is time consuming, and once the MDS analyses are carried out, the information would be redundant, since observations on how the spatial representations of sounds changed would be far more detailed than a number given by a statistical output. After all, the MDS is the tool chosen for the systematic analysis of the data matrices, and at this point it is sufficient to note that there are no obvious problems in the responses.

Finally, more than any statistical tests, the good interpretability of the overall results is taken as sufficient evidence for the reliability of the listeners. Indeed interpretability of data is preferred at all points of analysis over a single statistical value such as correlation coefficients. As we have seen in the review of Kewley-Port & Atal (1989) in §II.2.2, although given correlation values were high between the perceptual and auditory spaces, spatial configurations were capable of showing extra information about the relationship in that vowels were clustered according to their phonetic identity.

Now, we are ready to interpret the results. The results are presented separately for each stimulus set, and within a stimulus set, the results were structured to address the following questions in turn:

1. Do the stimuli spaces (of metric and nonmetric analyses) of a whole subject group give sensible configuration? Are they phonetically interpretable?
2. How large is the subject variability?
3. Are the stimulus spaces for the subject Groups A and B compatible?
4. How do the stimulus spaces change from the whole syllable set to the LPC4 set?

## 4.1 Whole syllable

### 4.1.0 Introduction

The ultimate goal of this section is to obtain a 'stable' and 'interpretable' perceptual dimension of whole fricative syllables. There is no short cut to reaching this goal, except to go about answering the four questions set out above, one by one.

Badness-of-fit was examined first to decide the dimensionality of the solutions. Once the dimensionality of the potential solution is decided, both interval and ordinal level stimulus spaces are presented and compared (§4.1.1). Although it is not discussed here, 2-way metric ALSCAL and INDSCAL analyses were also carried out to check the reliability of the results. Subject spaces and the split-half analysis of subject data on both interval and ordinal level solutions are checked for stability of the perceptual space of the group as whole. The perceptual dimensions were correlated with the phonetic properties of fricatives.

### 4.1.1 Stimulus space for the whole subject group

In order to determine the dimensionality of the solution, the changes in the badness-of-fit values with dimensionality are discussed first:

| | Dimensions | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| interval | .532 | .423 | .396 | / |
| ordinal | .404 | .214 | .136 | / |

**Table IV-1.** Badness-of-fit values, which show the fit error by interval and ordinal level proc-ALSCAL analyses, with the number of dimensions.

Since the analysis at the ordinal level concerns only the ordering of the objects in the space, and not with the differences in magnitudes, it provides better fit than the analysis at the interval level. Four-dimensional analysis was not permitted. It is clear that the elbows in the badness-of-fit occur after the second dimension for the interval analysis, and it is also very clear that the further dimensions improve the fit only moderately between the data and the model. For the ordinal analysis, either a two or three-

dimensional solution would be appropriate.

The two-dimensional interval and ordinal solutions are presented in the Figures IV-2 (a) and (b).

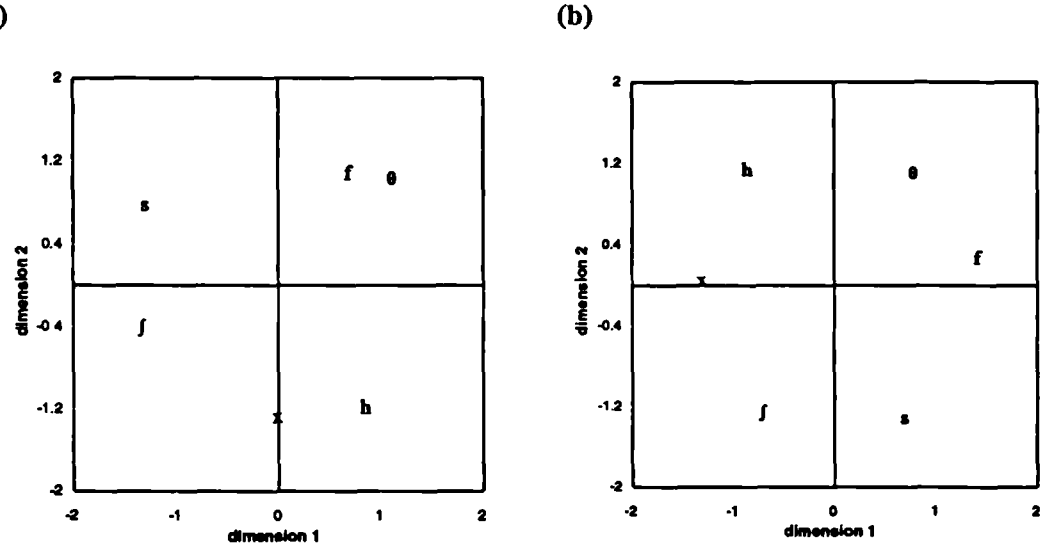



**Figure IV-2.** The stimulus configurations from (a) interval and (b) ordinal level solutions, for the whole syllable set.

At first sight, the configurations of interval and ordinal solutions look rather different. However, the interval dimension 1 is equivalent to the ordinal dimension 2, in that they make separation between the sibilants and nonsibilants. The actual orders on these dimensions look quite similar from each other, as shown below:

| interval 1 | ordinal 2 |
|---|---|
| θ | h |
| h | θ |
| f | f |
| x | x |
| ʃ | ʃ |
| s | s |

**Table IV-2.** Comparison of fricative ordering in the sibilance dimensions of interval and ordinal level analyses.

Also, the binary distinction between the sibilant versus nonsibilant fricatives is well maintained in both analyses, indicated by a horizontal line above, which divides the two groups of fricatives. The orders of fricatives on dimensions — interval dimension 2 and ordinal dimension 1 — are identical, and these dimensions place the fricatives according to their place of constriction, except for the fricatives /x/ and /h/. In any case, these two fricatives are placed extremely close to each other.

Figure IV-2 (c) shows a plot of dimension 3 against dimension 1 of the ordinal solution. The third dimension can only be interpreted as separating the fricatives which were closer together in earlier dimensions. The same effect was observed in the pilot experiments (§III.2.5.2). As it was the case previously, this dimension is not discussed any further.

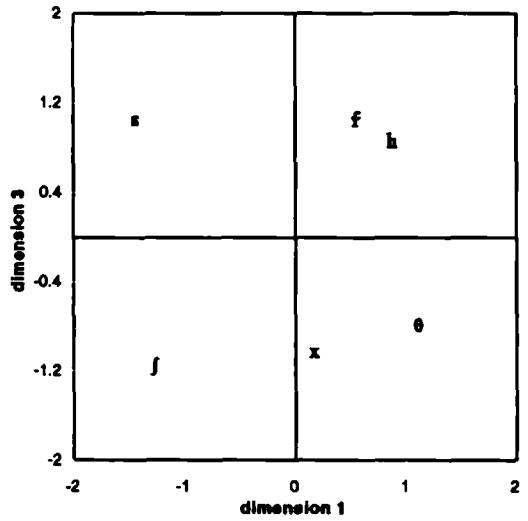



**Figure IV-2 (c).** The stimulus configurations of dimension 3 against dimension 1 of the ordinal level solution, for the whole syllable set.

We must just mention that the results of INDSCAL and nonmetric ALSCAL in pilot experiments were more similar to each other than observed here. A possible explanation is that, for the present data, the nonmetric solution is "compressing" the differences between the stimuli. This effect can also be seen in the subject space, which is discussed in the next section.

Overall, both analyses provide similar general stimulus spaces.

### 4.1.2 Subject space

The plots of subject spaces for fricative stimulus spaces are given in Figures IV-3 (a) and (b). The subject weights in the interval solution (Figure IV-3 (a)) show that most of the subjects lie on the quarter circle of the subject space, thus the MDS perceptual dimensions shown in the previous section are salient for most of the subjects. Subject 3 seems to be an outlier[4]. It also shows that the first group of subjects (1 to 10) gives more weight to the second dimension than to the first, except for subject 2. The second group of subjects (11 to 20) gives more emphasis to the first dimension than the second. Subject group differences are investigated further by the 'split-half' analysis in the next section.

(a) (b)

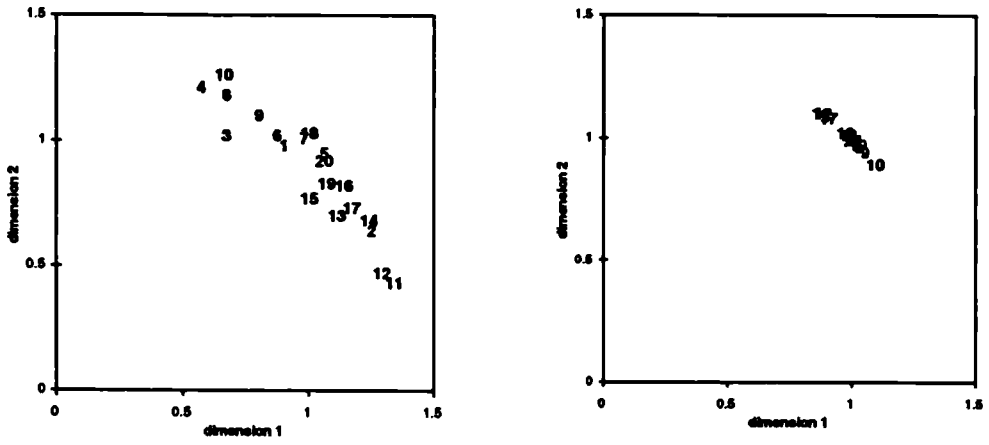



**Figure IV-3.** Dimension coefficients of the stimulus set showing the subjects' weight on each dimension, for (a) interval and (b) ordinal level solutions, in the whole syllable set.

This variation in subject weights has disappeared in the nonmetric analyses (Figure IV-3 (b)). All the subjects give almost equal weight to dimensions 1 and 2, with little variation. A summary table of average weights for each group and each analysis type is given below:

[4]A separate WMDS analysis, omitting subject 3, made little difference to the stimulus space configuration in Figures IV-2 (a) and (b). Thus WMDS seems to be able to account for such noise in the data.

| groups | dimension 1 | dimension 2 |
|---|---|---|
| A interval | 0.853 | 1.037 |
| B interval | 1.152 | 0.738 |
| A ordinal | 1.013 | 0.984 |
| B ordinal | 1.045 | 0.953 |

**Table IV-3.** Comparison of average subject weights in Groups A and B respect to both interval and ordinal level analyses for each perceptual dimension.

It was reported that "the INDSCAL weights often indicate greater relative importance between dimensions for a subject than do the ALSCAL weights" (Schiffman *et al.*, 1981: p244). As the interval option of ALSCAL is related to the INDSCAL analyses, this explains the discrepancies observed in the subjects' weights between the interval and ordinal level options. A comparison table showing two of the actual subject weights for the interval and ordinal levels illustrates the point further:

| subject | dimension 1 | dimension 2 |
|---|---|---|
| 3 interval | 0.68 | 1.02 |
| 3 ordinal | 1.00 | 1.00 |
| 19 interval | 1.08 | 0.83 |
| 19 ordinal | 1.00 | 1.00 |

**Table IV-4.** Comparison of individual subject weights in interval and ordinal level analyses for Subjects 3 and 19, for each perceptual dimension.

Overall, the individual subject weights for the two-dimensional interval and ordinal solutions indicate that both Groups A and B are adequately represented by the group stimulus spaces shown in Figures IV-3 (a) and (b). However, the actual comparison of the stimulus configuration of the two groups is carried out in the next section, to ensure the compatibility of the data from each of the two Groups A and B.

### 4.1.3 Group A vs. Group B

Because of the nature of the data collection method, the question of the compatibility of the data from the two separate listening Groups (A & B) needs to be addressed. Fox

(1995) implemented a **split-half** analysis method, which compares stimulus coordinates of a particular dimension between results from two split-halves of subjects, in order to test the reliability of a perceptual dimension. It was viewed that, if the corresponding dimensions of two split-halves of listeners showed any significant correlation, "a perceptual dimension was considered reliable, and thus likely to represent a psychologically real perceptual dimension for a group as a whole" (p2544).

Such an analysis was implemented here in order to check:

i) the stability of the group stimulus space;

ii) whether the individual stimuli spaces of Groups A and B are the same.

Weighted MDS analyses were carried out on the two subject groups separately and the configurations from the two groups are compared.

Both interval and ordinal solutions of the two groups are shown in Figures IV-4 to IV-5 respectively. The interval level solution is described first.

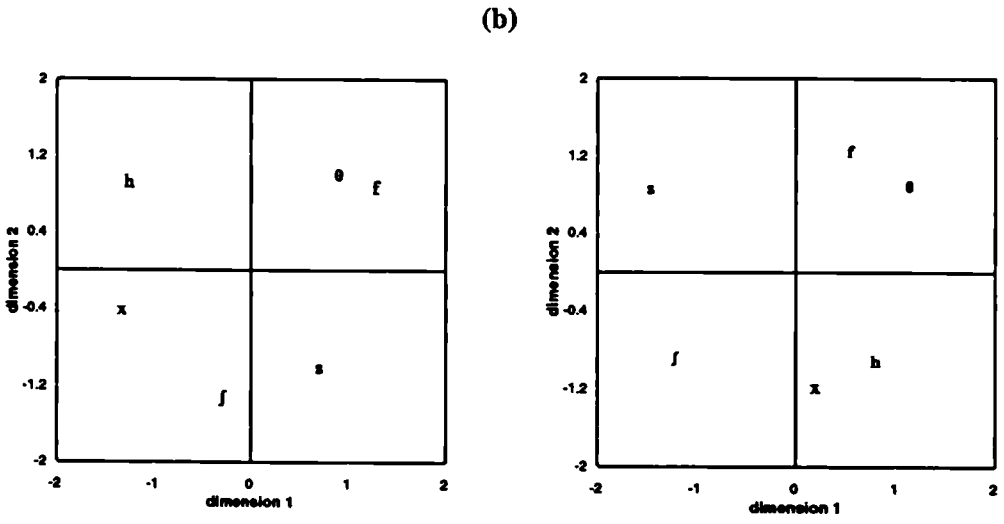



**Figure IV-4.** Stimulus spaces from interval level analysis (a) for Group A (b) for Group B, in the whole syllable set.

Before assessing how well the configurations of two different groups correspond to each other, firstly it can be noted that the dimension 1 of Group A separates places of articulation while the dimension 1 of Group B identifies sibilance. The significance of the orientation of each space can be judged by the average subject weights for each

dimension:

| | dimension 1 | dimension 2 |
|---|---|---|
| group A | 1.01 | 0.876 |
| group B | 1.144 | 0.759 |

**Table IV-5.** Comparison of average subject weights in Groups A and B for each perceptual dimension (interval level solution).

The average dimension weights indicate that the orientation of the axes cannot be ignored. Thus, comparing the dimension 1 of Group A and dimension 2 of Group B, we can see that, although the exact position of the fricatives differs, the fricatives are clearly ordered according to their place of articulation, except that the ordering was reversed for /x/ and /h/. In addition, we can see from the average subject weights that the place dimension is more important than the sibilance dimension for Group A whereas the reverse is true for Group B.

Although the axes are fixed, to give a quantitative *indication* of how well the two spaces of the two separate groups relate to each other, the method of canonical correlation analysis is applied. The results of such an analysis show that the data sets from the two different listening groups A and B provide spaces that are very highly correlated; 0.984 for the 'place' dimension and 0.946 for the 'sibilance' dimension.

For the nonmetric solutions, subject weights for each dimension are almost equal. In other words, the orientation of the axes was not fixed; rather, they could be rotated to obtain maximum correspondence. We can apply canonical correlation to determine how well the two spaces of the two separate groups relate to each other. The correlation coefficients were 0.998 for dimension 1 and 0.995 for dimension 2. The significance was 0.0001 and 0.0004 respectively, though the true significance is actually less than this, because of the lack of independence of the stimulus coordinates. The rotated nonmetric spaces of both groups are shown in Figures IV-5 (a) and (b).

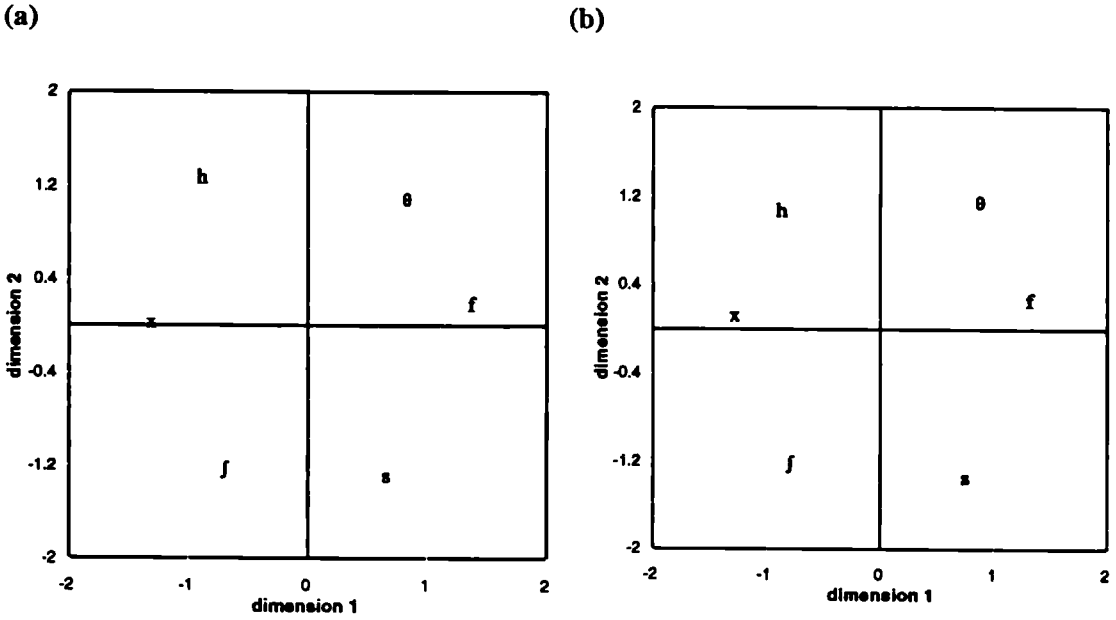



**Figure IV-5.** Stimulus spaces from ordinal level analysis (a) for Group A (b) for Group B, in the whole syllable set.

In keeping with the high values of the correlation coefficients, the configurations show close agreement.

### 4.1.4 Summary

i) For this set the whole group stimulus spaces were stable and the data for Groups A and B need not be interpreted separately.

ii) The nonmetric solution showed a 'compressing' effect on the differences between stimuli and subjects, thus it was less interpretable than the metric solution.

iii) The two perceptual dimensions were related to 'place' and 'sibilance'.

## 4.2 No-transition

### 4.2.0 Introduction

The aim in this section is to observe the effect of removing the transition sections in the perception of fricative syllables. Exactly the same procedures as in the previous set were applied to establish a suitable perceptual map. Whether the same relationship between interval and ordinal level solutions holds for this stimulus set is of particular interest here.

### 4.2.1 Stimulus space for the whole subject group

A comparison of the badness-of-fit values between interval and ordinal level analyses for one- through to four-dimensional solutions is as follows:

| | Dimensions | | | |
|---|---|---|---|---|
| | 1 | 2 | 3 | 4 |
| interval | .505 | .414 | .390 | .372 |
| ordinal | .399 | .219 | .137 | .088 |

**Table IV-6.** Badness-of-fit values, which show the fit error by interval and ordinal level proc-ALSCAL analyses, with the number of dimensions (no-transition set).

While the badness-of-fit decreases with increasing dimensionality for both analyses, the improvement is only moderate at the interval level after the second dimension. Because of this elbow effect after dimension 2, together with comparison reasons with the results of the previous set, only two-dimensional solutions are therefore interpreted here.

The combined group stimulus spaces from interval and ordinal level analyses are given below:

(a) (b)

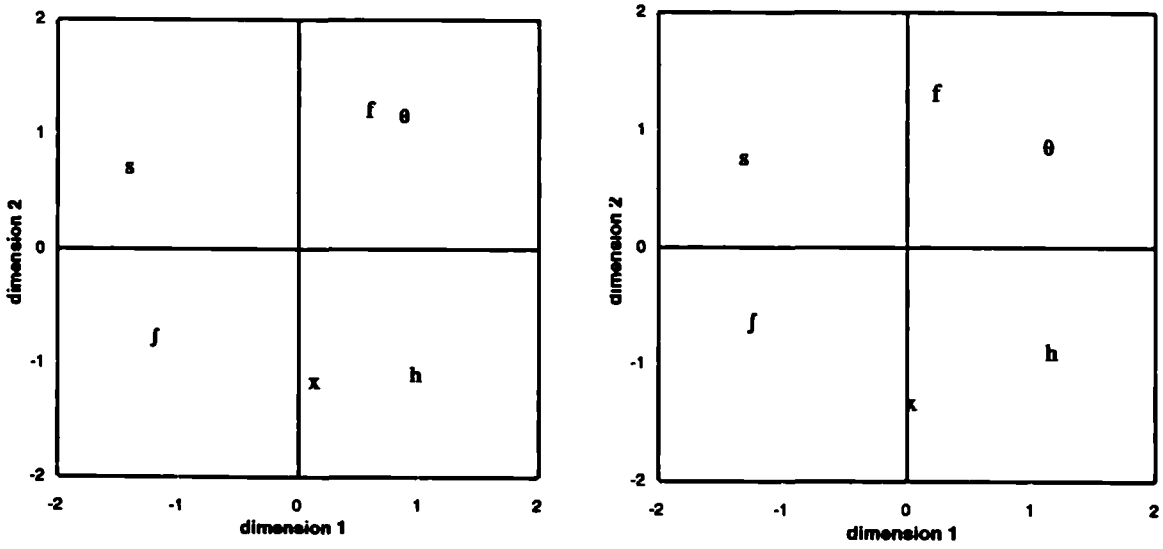



**Figure IV-6.** The stimulus configurations from (a) interval (b) ordinal level solutions for the no-transition set.

Even though the badness-of-fit values are miserable for the interval level analysis, the ordering of fricatives on stimulus spaces is exactly the same for both analyses. Dimension 1 clearly separates sibilance from nonsibilance. Dimension 2 orders fricatives according to their place of articulation, except that this order has been reversed for /x/ and /h/[5].

Therefore, the group stimulus spaces are extremely similar to those of the previous set.

### 4.2.2 Subject space

Subject spaces for both interval and ordinal level analyses are given in Figures IV-7 (a) and (b). As in the previous set, subject weightings of the first two dimensions are very high. Again, Subject 3 seems to be an outlier.

(a) (b)

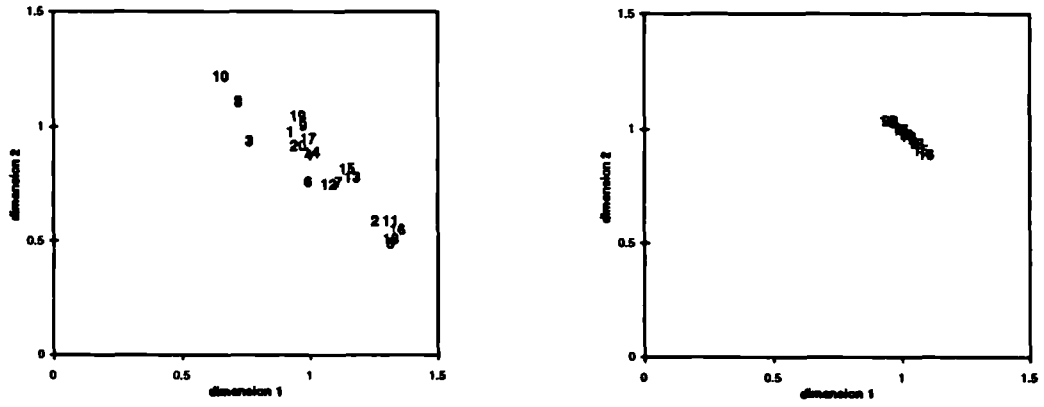



**Figure IV-7.** Dimension coefficients of the stimulus set showing the subjects' weights on each dimensions, for (a) interval, and (b) ordinal level solutions, in the no-transition set.

Most of the subjects gave more or less the same weight to both dimensions 1 and 2, for the metric solution. This is different from the results of the previous set, in which Group A gave more emphasis to dimension 2 and Group B gave more emphasis to dimension 1. In general, regardless of which group they belong to, subjects give more weight to

[5]This may be due to the fact that the quality of the velar fricative /x/ was such that it sounded more 'back' than it should have been, that is, more like a uvular fricative /χ/. This may have caused listeners to confuse the ordering of this fricative on the place dimension.

dimension 1. This observation is illustrated by the average weights for each group, as shown below:

| groups | dimension 1 | dimension 2 |
|---|---|---|
| A interval | 0.977 | 0.874 |
| B interval | 1.132 | 0.781 |

**Table IV-7.** Comparison of separate average subject weights in Groups A and B for each perceptual dimension (interval level solution).

For the nonmetric solution, as it was for the 'whole syllable' set, subjects show very little variation, as shown in Figure IV-7 (b).

Therefore, the results of subject spaces are satisfactory, and as far as subject weights are concerned, the Groups A and B can be analysed together. Separate MDS solutions of the two subject groups are compared in the following.

### 4.2.3 Group A vs. Group B

In order to examine the orientation of axes, the average subject weights for each dimension and each group are presented below :

| groups | dimension 1 | dimension 2 |
|---|---|---|
| A interval | 1.005 | 0.856 |
| B interval | 1.127 | 0.777 |

**Table IV-8.** Comparison of average subject weights in Groups A and B for each perceptual dimension (interval level solution).

It is clear from the average values of the subject weights that dimension 1 is given more emphasis and the orientations are fixed.

A comparison of the spatial arrangements of the two groups is shown in Figures IV-8 (a) and (b).

(a) (b)

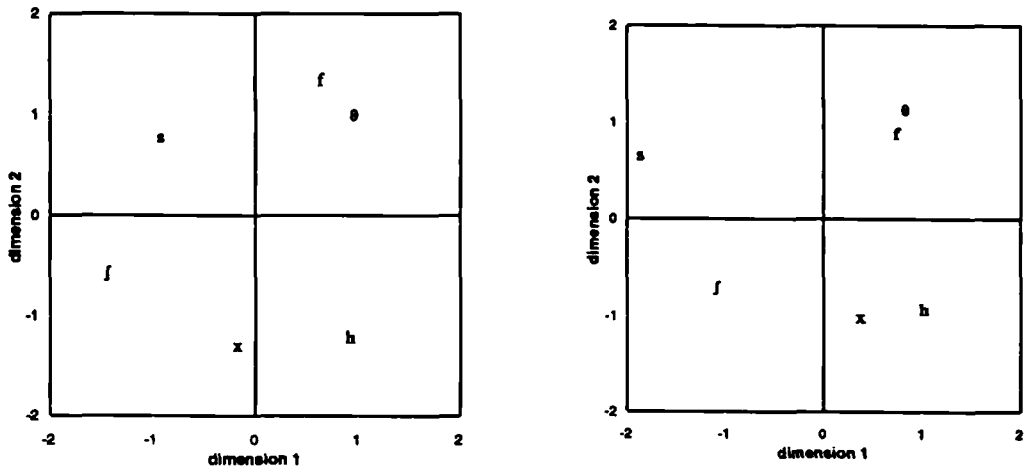



**Figure IV-8.** Stimulus spaces from interval level analyses of the no-transition set (a) for Group A (b) for Group B.

Although the exact locations of the fricatives on these two group stimulus spaces are different, dimension 1 nevertheless separates sibilants from nonsibilants, and dimension 2 orders fricatives according to their 'place', except that for Group B, the ordering of /f/ and /θ/ has been reversed. In any case, these two fricatives are extremely close to each other on dimension 2, Figure IV-8 (b). The canonical correlations were 0.989 for the 'place' dimension, and 0.934 for the 'sibilance' dimension.

At the ordinal level, the orientation of axes was not fixed, and the canonical correlation coefficients were 0.997 for dimension 1 and 0.833 for dimension 2. The poor agreement of dimension 2 is mainly due to the positions for /f/ and /θ/, as shown in Figures IV-9 (a) and (b), which are reversed.

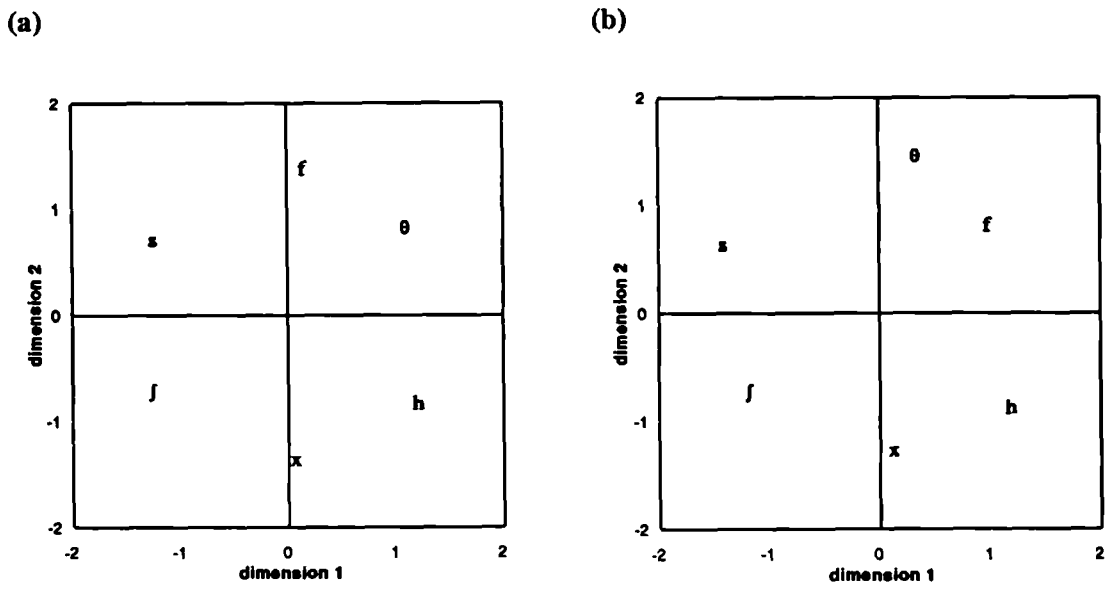



**Figure IV-9.** Stimulus spaces from ordinal level analyses of the no-transition set for (a) Group A (b) Group B.

Thus, the data sets from the two listening Groups A and B are closely correlated, thus the group stimulus space discussed in Figure IV-6 is very stable.

### 4.2.4 Summary

i) The stimulus spaces for the whole group in Figures IV-6 (a) and (b) are stable.

ii) The results are almost identical to those of the previous set. The canonical correlations between the interval level solutions of the previous set and the present set were 0.999 and 0.995. They were both significant. Thus, transition sections are not important in fricative perception, at least when the following vowels are identical.

iii) Note that, so far, the rank order between the interval and ordinal data has been identical for the 'place' dimension. This means that, although the basic arrangement of the stimuli is the same for both analyses, at the ordinal level, the actual clustering of the fricatives (which are shown close in their place of articulation on the interval solution) is not shown. This observation is in line with the warning by Schiffman *et al.* (1981) that "Relaxation of the interval measurement level assumption improves *fit*, but may worsen *interpretability*, ..." (p251) [my italics].

This provides justification for not interpreting the nonmetric solutions for the remainder of the stimulus sets, unless the results are otherwise.

iv) The separate sections discussing subject space are omitted from now, although the results are mentioned where necessary. This is because the results were very similar for all the stimulus sets, indicating that the subjects behaved consistently throughout the tests.

## 4.3 Cut-out

### 4.3.0 Introduction

The aim of this section is to establish the statistical validity of the perceptual map, based on the fricative section only, which has already been investigated with five listeners in the preliminary analyses (in Chapter III). Same analysis procedures are adopted as in the two previous sets, except that only interval level solutions are discussed here.

### 4.3.1 Group stimulus space

Changes in the badness-of-fit values with increasing dimensionality were similar to the previous two sets, thus two dimensional solutions are thought to be adequate for representing the data.

The two-dimensional group stimulus spaces from interval level analysis are presented in Figure IV-10.

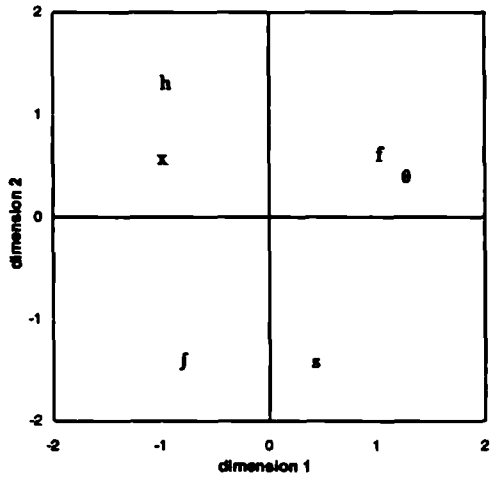



**Figure IV-10.** The group stimulus spaces from interval level analysis, for the 'cut-out' set.

As in the two previous sets, the 'place' and 'sibilance' dimensions are clearly maintained. The only difference observed in the group stimulus space from the previous sets is that the ordering of /f/ and /θ/ was reversed. This was already true for the Group B in the previous set (Figure IV-8 (b)). In any case, they are placed very close to each other[6].

In the preliminary analysis, the place dimension was clear, but the sibilance dimension was not shown. This is different from the result shown here. Although the stimuli were normalised in terms of their duration and overall RMS level, the sibilance distinction is still clearly maintained.

### 4.3.2 Group A vs. Group B

Separate MDS analyses on Groups A and B showed that the average subject weights of the first two dimensions are similar at interval level analysis, as shown below:

| groups | dim. 1 | dim. 2 |
|---|---|---|
| A interval | 0.952 | 0.938 |
| B interval | 1.038 | 0.913 |

**Table IV-9.** Comparison of average subject weights in Groups A and B for each perceptual dimension (the cut-out set).

However, a plot of the subject space shows that there are clear differences in subject weights for each individual subject. This is presented in Figure IV-11.

[6]Furthermore, the phonetic distinction between these two fricatives is not well maintained in South Eastern dialects. To a certain extent, therefore, such confusion is anticipated.

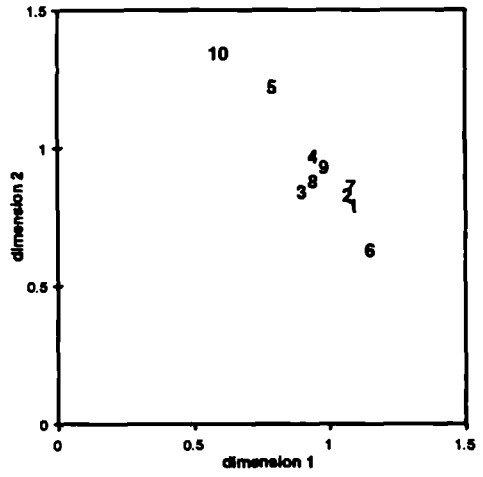



**Figure IV-11.** Subject space for interval level analysis for Group A only, in the cut-out set.

Thus in the stimulus spaces for each Group A and B, shown in Figures IV-12 (a) and (b), dimension 1 of Group A corresponds to dimension 2 of Group B; these are the 'place' dimensions. Although the actual ordering of /f-θ/ and /x-h/ in the two subject groups is reversed, the fricatives form three pairs /f-θ/, /s-ʃ/, and /x-h/, and they are organised in a similar way for both subject groups. The canonical correlation coefficients were 0.980 and 0.911.

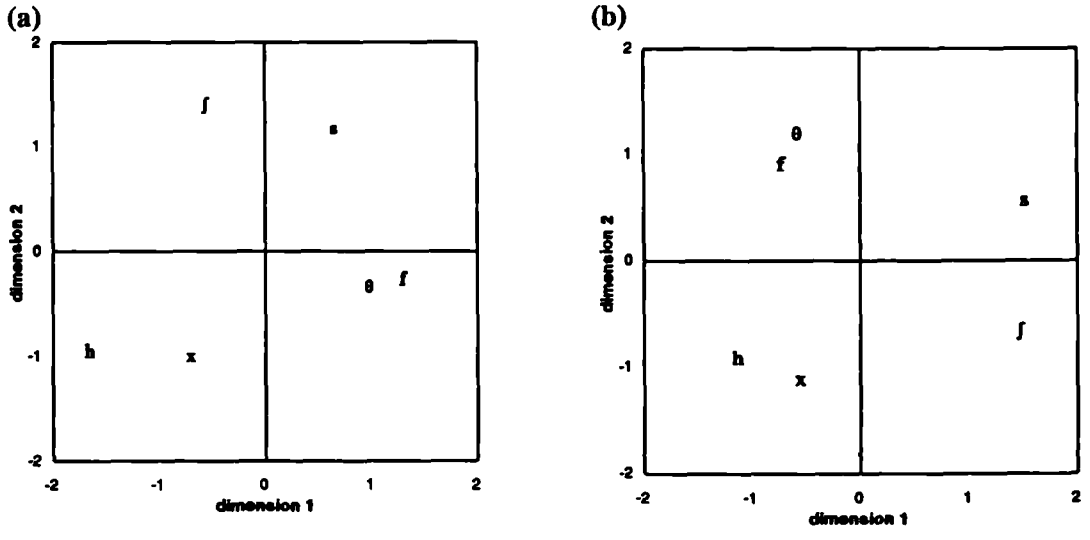



**Figure IV-12.** Stimulus spaces of the cut-out set from the interval level analysis for (a) Group A, and (b) Group B.

### 4.3.3 Discussion

i) The group space shown in §4.3.1 is stable, and the results from the two subject Groups A and B are compatible, but not as closely related as in the two preceding sets.

ii) Overall results are very similar to those of the previous set (correlation coefficients = 0.998, 0.974), and the perceptual dimensions correspond to 'place' and 'sibilance' of fricatives .

iii) From the results of the three stimulus sets so far, we have seen that two-dimensional solutions adequately represent the data, without interpreting any further dimensions. Since badness-of-fit values were more or less the same in the remainder of the stimuli sets, two-dimensional solutions are thought to be adequate for the remainder as well.

## 4.4 LPC22

### 4.4.0 Introduction

The purpose of this section is to investigate the effect of 'dynamic' acoustic properties of fricatives in perception. In this set, the average spectral shape of each fricative was modelled by LPC synthesis with 22 coefficients, and repeated to make up to 400 ms.

### 4.4.1 Group stimulus space

The two-dimensional solution from interval level analysis is shown in Figure IV-13.

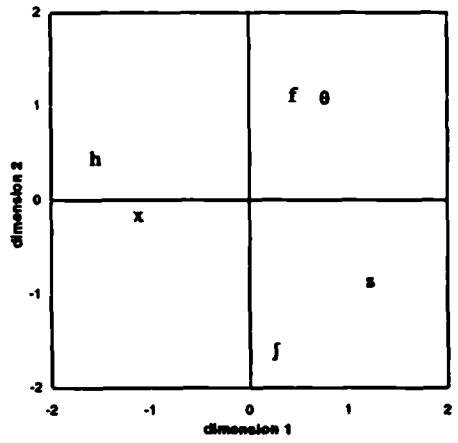



**Figure IV-13.** The two-dimensional interval level solution for the LPC22 set.

Dimension 2 can be labelled 'sibilance', but the ordering of fricatives on dimension 1 is not strictly according to 'place'. Although the group stimulus space was significant for all the subjects, their weightings on each dimension were different. Thus, the orientation of the space was fixed. However, if we could imagine that the map was turned clockwise by about 30 degrees, we find that the configurations are not that different from the previous set (Figure IV-10). Also, the pairings of fricatives /f-θ/, /s-ʃ/, and /x-h/ are maintained. However, the dimensions from the WMDS solution are meant to be directly interpretable, and this was not the case for this set. The stability of this solution is checked with a 'split-half' analysis on the Groups A and B in the following.

### 4.4.2 Group A vs. Group B

Separate WMDS analyses were carried out for Groups A and B. The group stimulus spaces from interval level solutions are presented in Figures IV-14 (a) and (b).

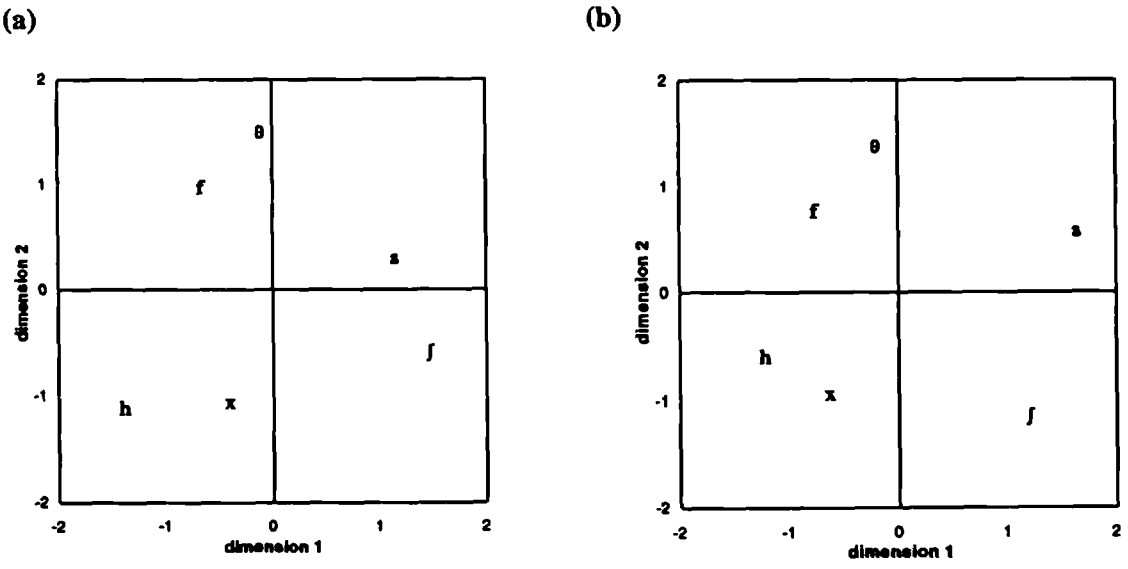



**Figure IV-14.** Group stimulus spaces from interval level analysis for (a) Group A, and (b) Group B, in the LPC22 set.

Although canonical correlation coefficients between the two spaces were higher than in the previous sets (0.997, 0.926), the interpretability of the spatial arrangements between the two subject groups is rather different; for Group A, sibilance and place dimensions are

apparent, as before. However, this was obviously not true for Group B, and also, the pairing of the fricatives is much more relaxed than in the previous set (in Figure IV-12 (a)). The incompatibility of Groups A and B is another indication that the results of this set are different from the three previous sets. Thus, the stability of the group stimulus space, in Figure IV-13, could not be established. Further split-half analyses of Group A and B data are appropriate here. Separate WMDS analyses on two split-halves of Groups A and B are conducted. Figures IV-15 (a) and (b) show the stimulus spaces for the first and second half of subjects in Group A.

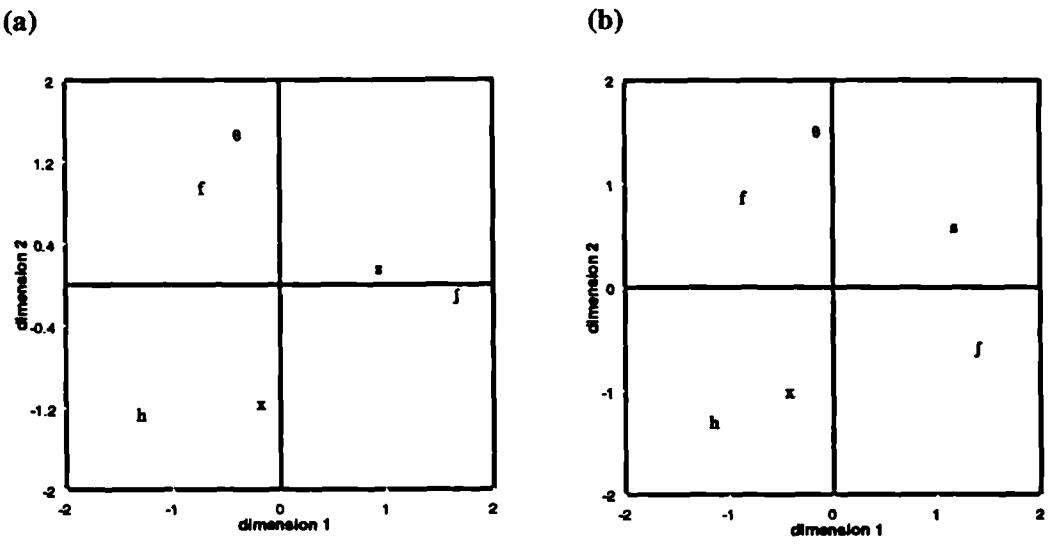



**Figure IV-15.** Group stimulus spaces for Group A; (a) subjects 1-5, and (b) subjects 6-10 , in the LPC22 set.

Although the exact location of fricatives is different on the two spaces, the patterns of perceptual arrangements are the same; dimension 1 clearly represents sibilance, and dimension 2 represents place.

The group stimulus spaces of subjects 11-15 and 16-20 are presented in Figures IV-16 (a) and (b).

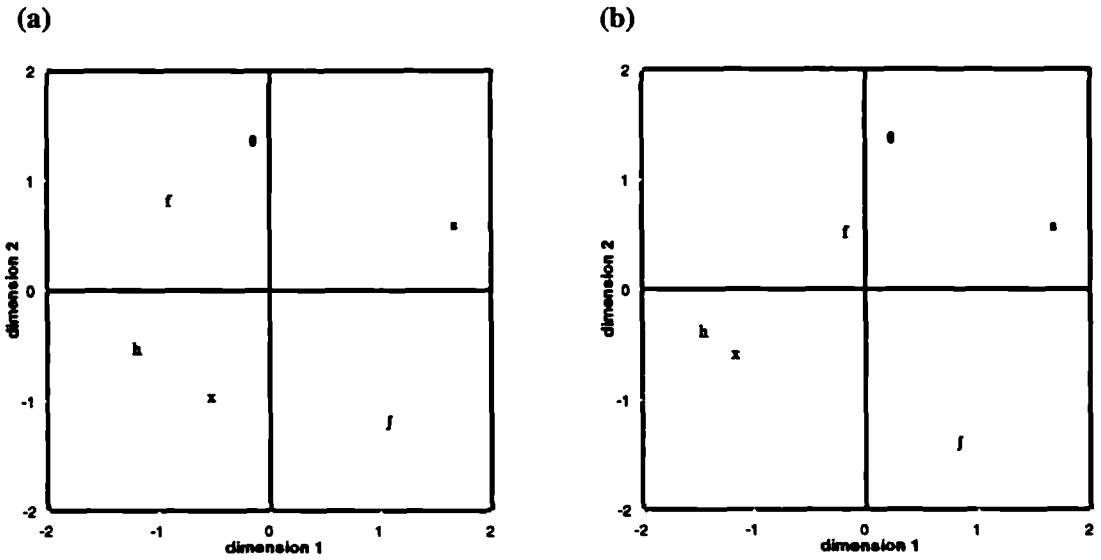



**Figure IV-16.** Group stimulus spaces for (a) the first half of Group B, and (b) the second half of Group B, in the LPC22 set.

The spatial representations from the two halves of Group B do not agree as well as in Group A. However, they share similar features, in that /f-θ/ and /x-h/ form loose pairings, and /s/ and /ʃ/ are rather far apart. Thus, the overall configurations are quite similar, and these features establish the results of Group B as being quite different from those of Group A.

### 4.4.3 Discussion

The group stimulus spaces for the LPC22 set should be interpreted separately for Groups A and B. The results for Group A were very similar to the previous sets (canonical correlations: 0.999, 0.979). The 'place' feature was confused for Group B (canonical correlation to cut-out set: 1.000, 0.962).

The difference in the results must be attributed to response mode. That is, for the subjects in Group B who had been listening to more synthetic versions of fricatives, we must assume that the static spectral information is not sufficient to trigger the speech mode of perception. On the other hand, for Group A, which had been tuning in to natural speech materials throughout the tests, static fricative spectral information was sufficient to allow subjects to perceive the fricatives according to their phonetic properties.

## 4.5 LPC10a

### 4.5.0 Introduction

The aim of this section is to observe the effect of the following /ɑ/ in the perception of the fricatives, in comparison with the next set, where the vowel was removed. The initial hypothesis was that the following vowel may help to trigger the speech mode of perception.

### 4.5.1 Group stimulus space

The two-dimensional solution from the interval level analysis is presented in Figure IV-17. Dimension 1 is clearly the 'sibilance' dimension. Strictly speaking, dimension 2 cannot be interpreted as 'place', since /s/ and /ʃ/, which are adjacent to each other in terms of place of articulation, are placed far apart. This seems to indicate that the group stimulus spaces are becoming less phonetically interpretable. The question of whether they are more auditorily related is addressed in the next chapter. Meanwhile, we must examine whether the stimulus space is stable, with split-half analysis on the two subject groups in the following section.

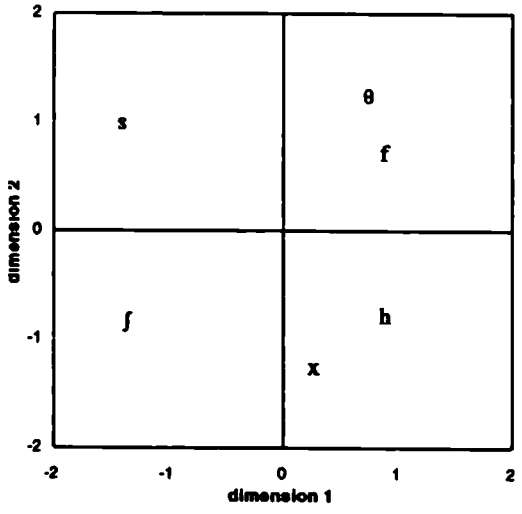



**Figure IV-17.** The two-dimensional interval level solution for the LPC10a set.

### 4.5.2 Group A vs. Group B

The two-dimensional solutions from the interval level analysis of Groups A and B are shown in Figures IV-18 (a) and (b). For Group A, the fricatives, /f θ/ and /x h/ are paired up on the space, but the /s-ʃ/ pairing is rather widely spaced. /s/ and /ʃ/ are even more apart in Figure IV-18 (b), and the ordering of fricatives according to their place of articulation fails to be maintained here. However, the sibilance dimension (dimension 1) is very distinctively shown in both spaces.

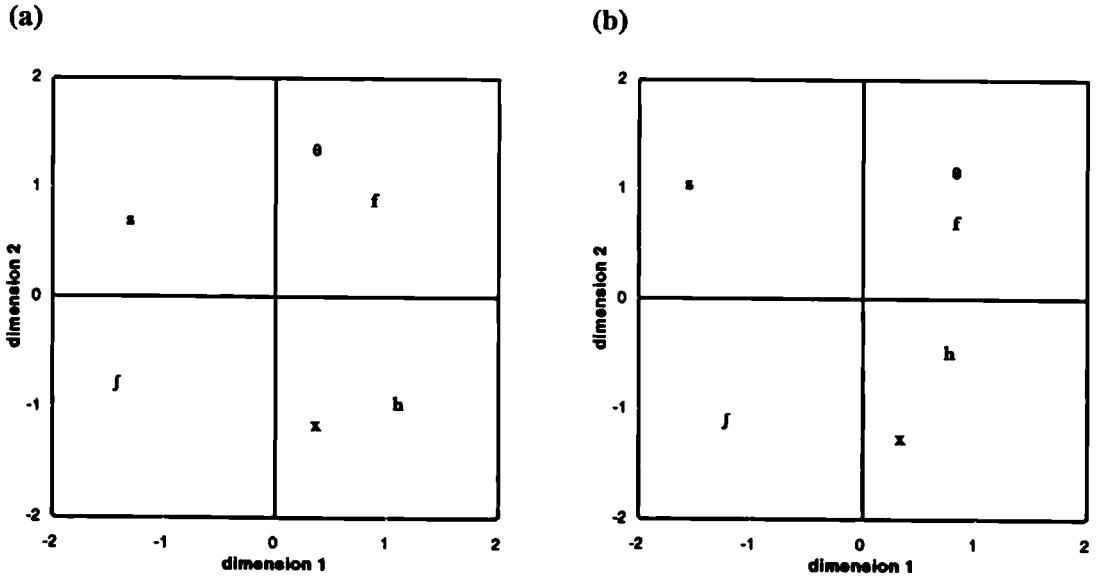



**Figure IV-18.** Stimulus spaces from interval level analysis for (a) Group A, and (b) for Group B, in the LPC10a set.

Since the results from the two subject groups are very different, further split-half analyses are carried out to check the stability of the configurations above. The stimulus spaces from the two halves of Group A are almost identical, as shown below.

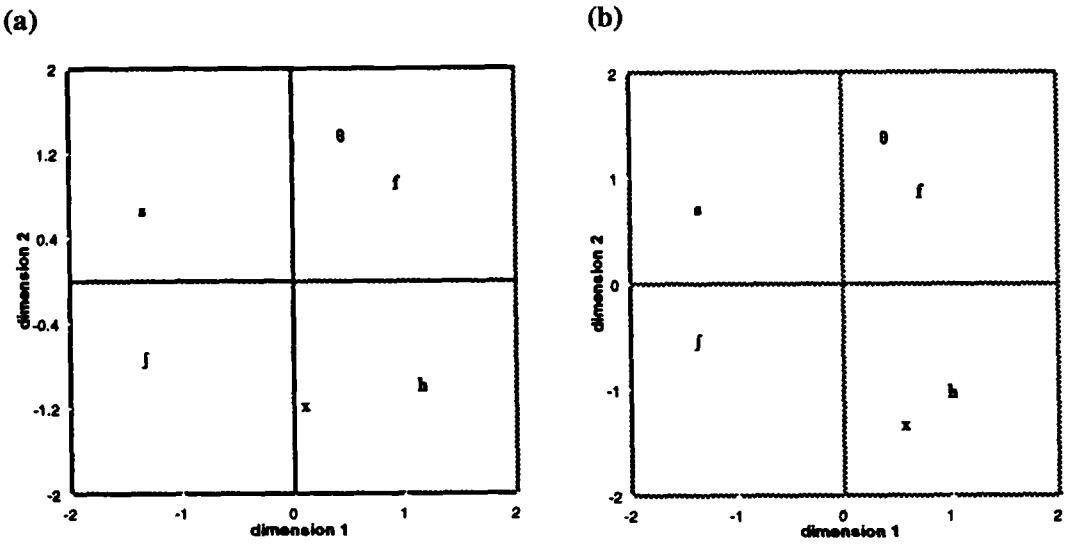



**Figure IV-19.** Stimulus spaces from the interval level analysis for Group A; (a) subjects 1-5, and (b) subjects 6-10, in the LPC10a set.

The 'place' and 'sibilance' dimensions are clearly shown, especially for (b).

The stimulus spaces from the two halves of Group B are shown in Figure IV-20 (a) and (b).

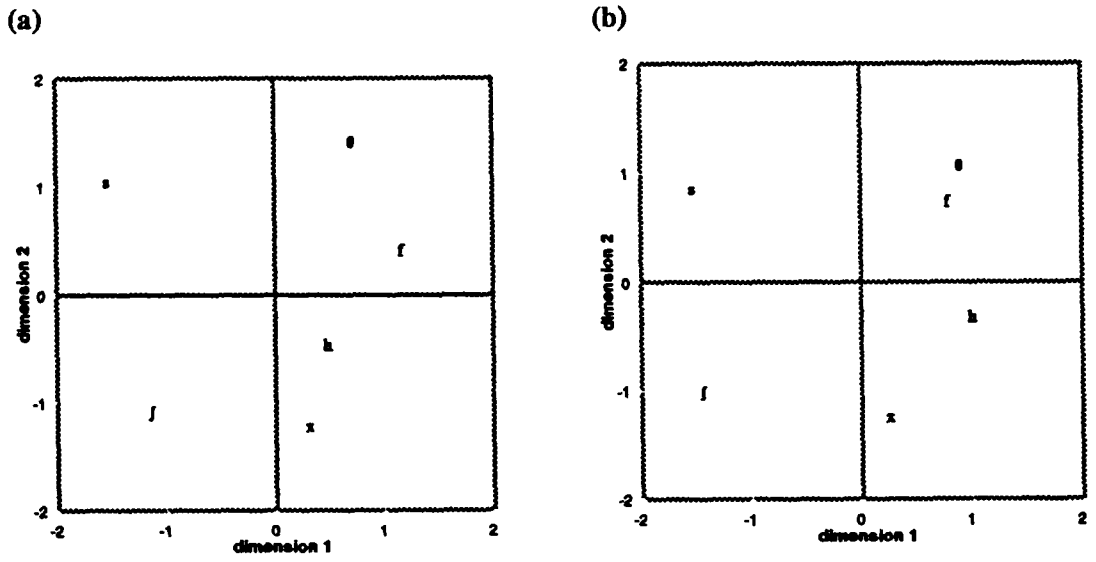



**Figure IV-20.** Stimulus spaces from the interval level analysis for Group B; (a) subjects 11-15, and (b) subjects 16-20, in the LPC10a set.

Although /f/ and /θ/ are more apart on the stimulus space for subjects 11-15, the basic organisation of the fricatives on the two split-halves of Group B is the same. Thus the stimulus space of Group B is also stable.

### 4.5.3 Summary

i) The stimulus space for the whole subject group was not stable, and data from Groups A and B need to be interpreted separately.

ii) For Group A, 'place' and 'sibilance' dimensions are still clearly separated, but for Group B, only the 'sibilance' dimension can be interpreted. The configurations from each split-half of the Groups A and B were almost identical.

iii) The following vowel seems to have failed to trigger the speech mode of perception for Group B. For Group A, the effect of a following vowel can only be clarified by comparison with the perceptual map of the LPC10 set.

For the analysis of the remaining two sets, results from Groups A and B are processed separately from the outset.

## 4.6 LPC10

In Chapter III, LPC10 fricatives without any following vowel were not perceived phonetically. Here, statistical validity is sought with 20 listeners. Also, the perceptual effect of the following /ɑ/ vowel in the previous set can be clarified by comparison.

The two-dimensional solution from the interval level analysis for Group A (Figure IV-21 (a)) shows that the pairing of the fricatives observed in the previous set (LPC10a) is clearly looser here. The stimulus spaces from the two split-halves of Group A, shown in Figures IV-22 (a) and (b), clearly demonstrate that the 'place' dimension is no longer interpretable in this set. Since the fricative spectral information was exactly the same as in the previous set, it seems that the following synthetic vowels in the previous set could help to trigger the speech mode of perception for Group A. Also, we could tentatively conclude that there is a perceptual switch from the LPC10a set to the LPC10 set for Group A. This observation can be only confirmed by examination of the corresponding auditory maps for these two sets. If there really is a perceptual switch, then the perceptual

map for LPC10 is expected to be much more highly correlated to the auditory map than is the perceptual map for LPC10a.

For Group B (Figure IV-21 (b)), no specific difference between the configurations of this set and the previous set is observed.

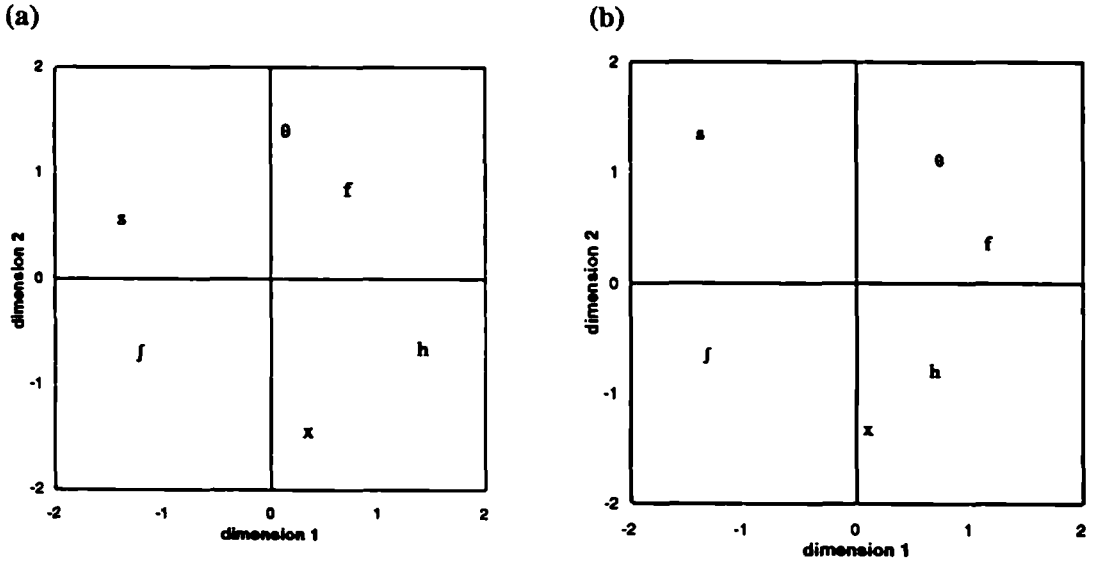



**Figure IV-21.** Stimulus spaces from the interval level analysis for (a) Group A, and (b) for Group B, in the LPC10 set.

Stimulus spaces from the two split-halves of Group B are presented in Figures IV-23 (a) and (b). Figure (a) looks a little different from (b), since /s/ and /ʃ/ are noticeably apart in (a). In fact, Figures IV-22 (a), (b) and Figure IV-23 (b) are very similar. Indeed, if the subjects in Group A have really 'switched' from the speech to nonspeech mode of perception, then there should be no differences between the stimulus spaces of Groups A and B. Any differences there are between Groups A and B may be clarified with reference to their auditory map in the next chapter.

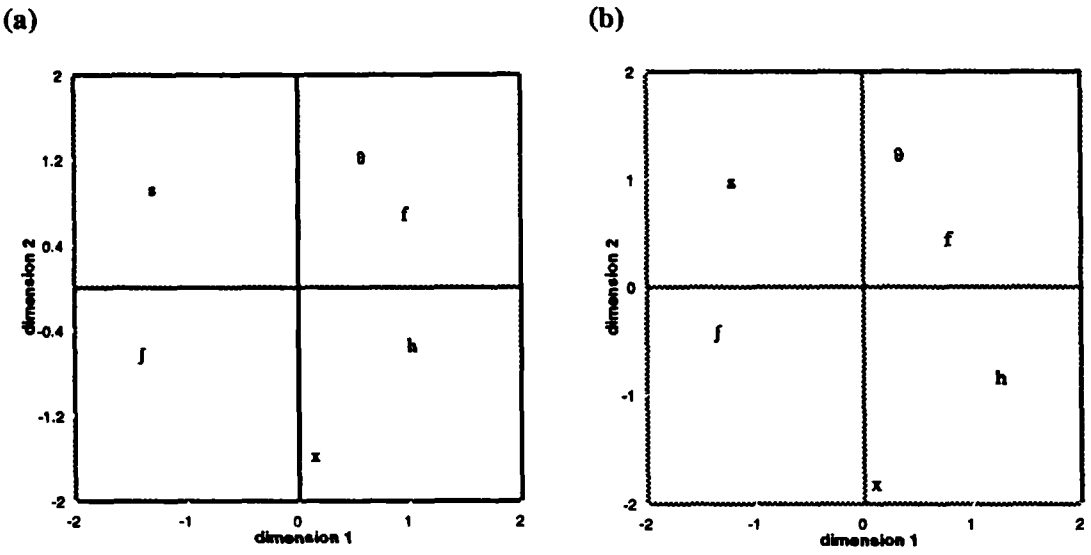



**Figure IV-22.** Stimulus spaces from the interval level analysis for Group A; (a) subjects 1-5, and (b) subjects 6-10, in the LPC10 set.

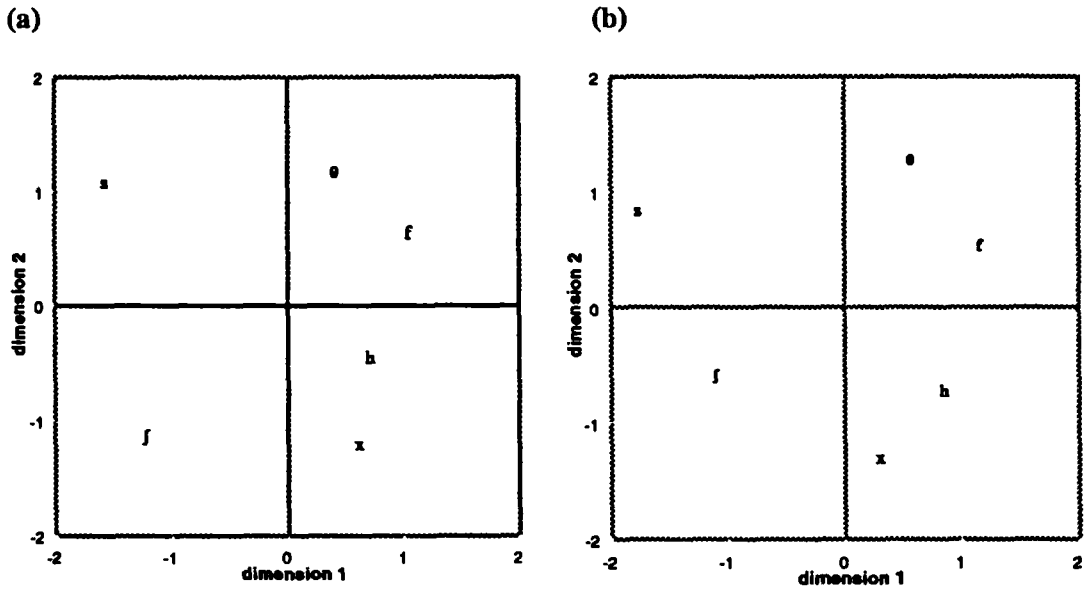



**Figure IV-23.** Stimulus spaces from the interval level analysis for Group B; (a) subjects 11-15, and (b) subjects 16-20, in the LPC10 set.

## 4.7 LPC4

In this set, we do not expect the perceptual maps to be phonetically interpretable. The stimuli in this set were fricatives modelled with two formants. Thus, in terms of acoustic form, the stimuli were similar to the two-formant white noises in §III.4. It was initially

hypothesised that the correlations between the perceptual and auditory maps for these speech-modelled spectra may not be as close as those found in the case of the white noises. This issue can only be fully addressed in the next chapter. Meanwhile, the aim in this section is to establish a reliable perceptual map which can be correlated with the auditory map in the next chapter.

The general patterns of the two-dimensional solutions at the interval level for Groups A and B are completely different (in Figure IV-24 (a) and (b)). We are already familiar with the perceptual pattern shown in (a) below, to the extent that dimension 1 indicates sibilance, and dimension 2 shows 'traces' of the place dimension. However, in (b), the fricatives /s-ʃ/ and /f-θ/ are far apart on dimension 2, and thus, phonetic interpretation is not possible.

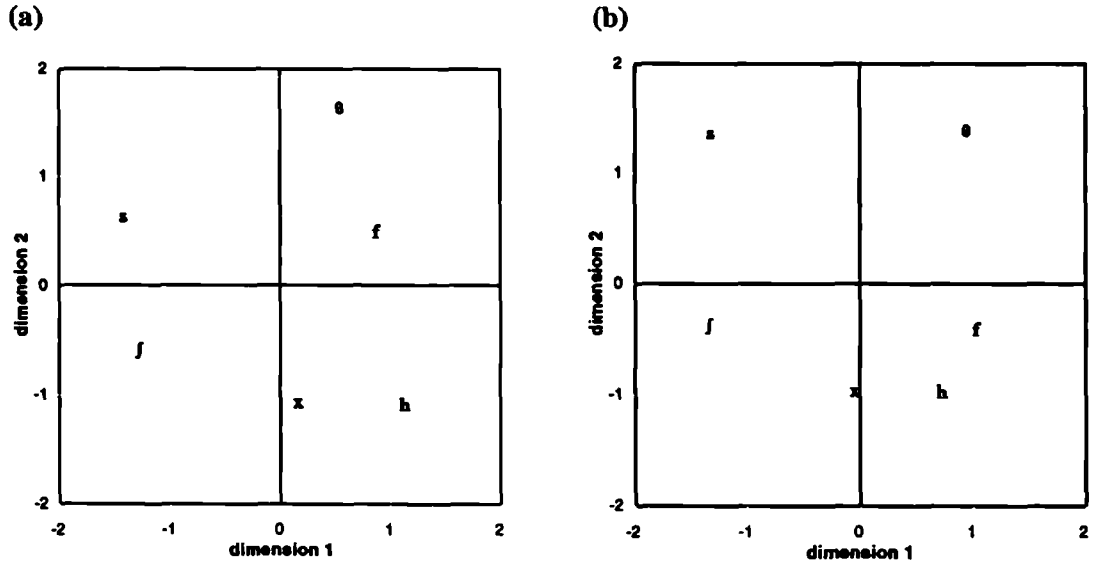



**Figure IV-24.** Stimulus spaces from the interval level analysis for (a) Group A, and (b) Group B, in the LPC4 set.

To complete the analyses, stimulus spaces of the split-halves of each subject group are shown in Figures IV-25 and IV-26. Figures IV-26 (a) and (b) are almost identical to each other. Thus, the group stimulus space for Group B is very stable. However, Figures (a) and (b) in IV-25 for Group A are slightly different, but not in any significant way. In fact, the plot in (b) displays roughly the same basic organisation as (a), although /f/ and /θ/ are

further apart. A possible explanation for this is that, while the stimuli are very artificial for this set, some subjects in Group A may be able to make a mental reference to the original fricatives. The same reason can apply to the differences between the spaces for Groups A and B. However, the perceptual strategies of these subjects can only be clarified by observing the correlations with the auditory map.

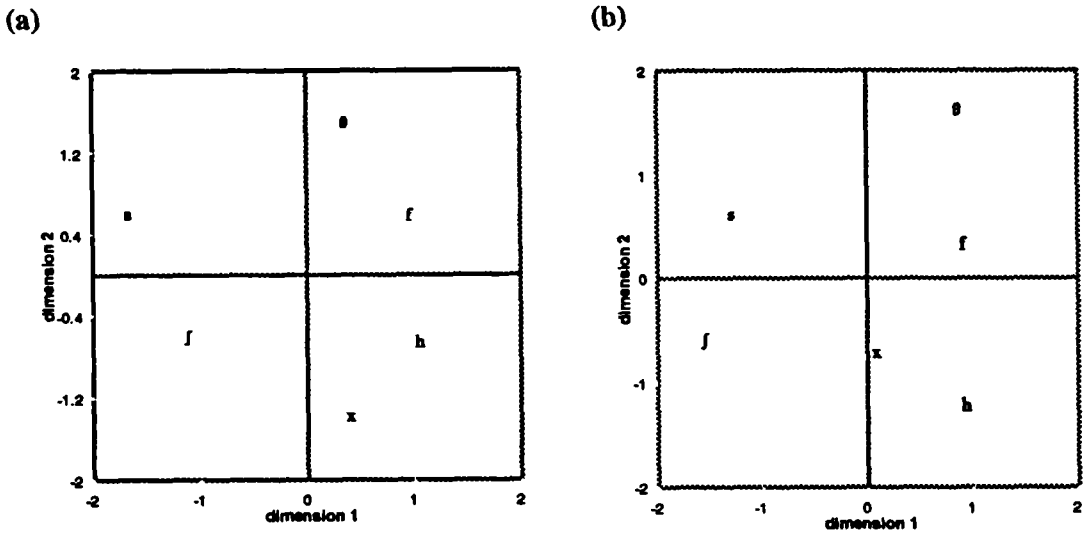



**Figure IV-25.** Stimulus spaces from the interval level analysis for Group A; (a) subjects 1-5, and (b) subjects 6-10, in the LPC4 set.

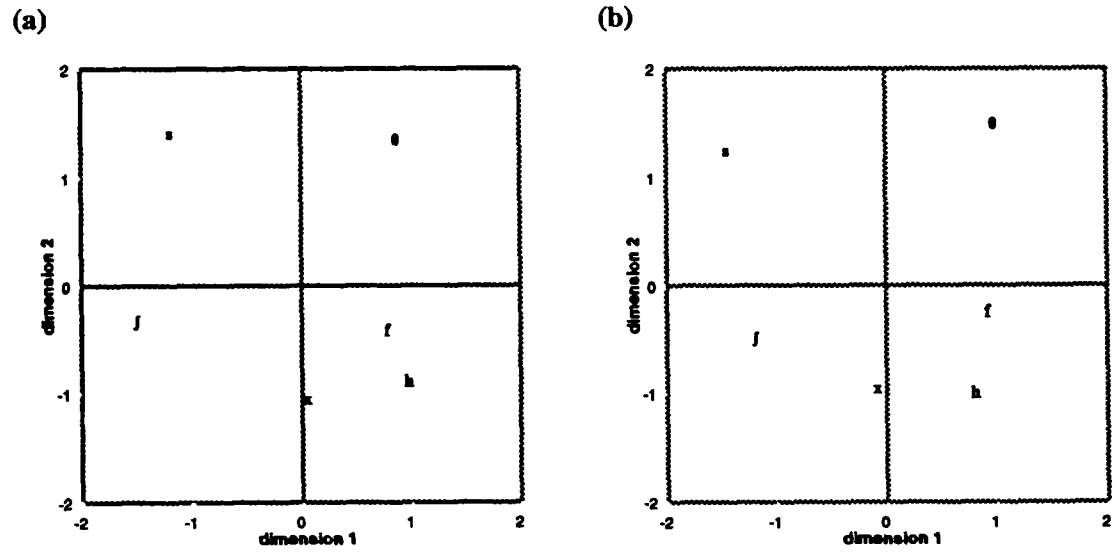



**Figure IV-26.** Stimulus spaces from the interval level analysis for Group B; (a) subjects 11-15, and (b) subjects 16-20, in the LPC4 set.

## 4.8 Overview of the results

The phonetic interpretability of solutions for the whole group decreased quite noticeably from LPC22 set downward. For the sets referred to as, 'whole syllable', 'no-transition', and 'cut-out', the two perceptual dimensions from MDS were clearly related to 'sibilance' and the 'place' properties of fricatives. While the 'sibilance' dimension was maintained throughout, the place dimension was less easily distinguished in the stimulus space of joint groups for the sets LPC22 to LPC4.

The compatibility of the data between the two subject groups A and B also changed from the LPC22 set downward. Although there were no clear differences between the Groups in terms of the subject weights, a comparison of the actual stimulus configurations from split-half analyses showed that results from the two groups were similar only for the whole syllable set, the no-transition set, and the cut-out set. For Group A, both 'place' and 'sibilance' dimensions were maintained in the sets LPC22 and LPC10a, but in the case of the sets LPC10 and LPC4, the place of articulation dimension was obscured. For Group B, the place dimension was unclear from LPC22 to LPC4. A summary table is given below:

| | Group A | | Group B | |
|---|---|---|---|---|
| | sibilance | place | sibilance | place |
| LPC22 | ✔ | ✔ | ✔ | ✗ |
| LPC10a | ✔ | ✔ | ✔ | ✗ |
| LPC10 | ✔ | ✗ | ✔ | ✗ |
| LPC4 | ✔ | ✗ | ✔ | ✗ |

**Table IV-10.** A summary of phonetic interpretability of the MDS dimensions in the sets LPC22, LPC10a, LPC10, and LPC4, for Groups A and B.

Correlation analyses between the whole syllable set and the remaining sets with respect to the coordinates for the place dimension may support this observation. For the no-transition and cut-out sets, joint group data are employed in the correlation analyses, as shown below:

| Compared sets | correl | R-sq. |
|---|---|---|
| No-tran. | .985 | .971 |
| Cut-out | .950 | .903 |

**Table IV-11 (a).** Correlations between the whole syllable, no-transition, and cut-out sets of the place dimension in their group stimulus spaces.

As observed spatially (compare Figure IV-2 (a) with Figures IV-6 (a) and IV-10), correlations with the whole syllable set are rather high for these sets.

In the case of the remaining sets, Groups A and B were treated separately:

| | Group A | | Group B | |
|---|---|---|---|---|
| Compared sets | correl | R-sq | correl | R-sq |
| LPC22 | .955 | .911 | .889 | .791 |
| LPC10a | .971 | .942 | .899 | .809 |
| LPC10 | .948 | .899 | .912 | .832 |
| LPC4 | .932 | .869 | .765 | .585 |

**Table IV-11 (b).** Correlations between the whole syllable set and the other sets with respect to the 'place' dimension; Groups A and B are presented separately.

For Group A, the correlations are high for the sets LPC22 and LPC10a. Although the correlation for the LPC10 set is not exceptionally low (compared to the correlation for the cut-out set, for instance, which was .950), nevertheless the correlation is no higher than for the earlier sets, and the place dimensions on the split-halves of Group A for LPC10 (shown in Figures IV-22 (a) and (b)) were noticeably obscured. Thus, it may be concluded that the correlation values do support the spatial observations. For Group B, the correlations are clearly not high from the sets LPC22 to LPC4.

These differences in the subject Groups A and B may be attributed to the different modes of perception. The same difference in the modes of perception is also a possible explanation for the fact that the /ɑ/ vowel in the LPC10a set produced more 'phonetically' based judgments from the listeners in Group A than from those in B.

Although the correlations between whole syllable stimuli and other stimuli for the place dimension support spatial interpretations, canonical correlation analyses of the

stimulus spaces for adjacent sets show that, in fact, the spatial configurations as a whole are subject to only a very slight change between one stimulus set and another. The coefficients range from 1 to 0.937, and these were all significant.

Returning to the question raised in §III.5, as to whether there might be a perceptual 'switch' or a series of 'gradual changes' from natural to synthetic speech, the response must remain indeterminate at this stage. According to the canonical correlation values between the stimulus spaces of different sets, the change is gradual from one set to the next. However, according to the phonetic interpretability of MDS configurations, there is a perceptual 'switch' from the cut-out set to LPC22 set for Group B, while this switch is observable from LPC10a to LPC10 for Group A.

If the perceptual configurations of those sets which had poor phonetic interpretability, were to show a better correlation with the corresponding auditory configurations, then the initial hypothesis would be confirmed. However, since the changes between the sets are very small, any differences in the correlations with the auditory spaces may be also very small. This issue is investigated in the next chapter.

# *Chapter V. The Relationship between perceptual and auditory spaces*

## 1 Introduction and objectives

In this chapter, we will implement various auditory distance models in an attempt to account for the perceptual results obtained in the last chapter. One of the principal objectives in this study is to test how well each auditory model predicts the perceptual similarity of the fricatives or fricative-like sounds, and how the predictive power of each model varies according to the degree of naturalness in the stimuli.

In Chapter IV, the similarity judgements on fricative pairs were represented as non-linear functions of distances in MDS spaces. The stimulus sets ranged from naturally produced fricative syllables to fricative-like sounds featuring varying degrees of artificiality in the signal. It was found that the stimuli involving natural fricatives tended to have MDS dimensions which were readily interpretable in terms of known phonetic properties (features). However, as the stimuli became progressively more artificial, this phonetic association became gradually less apparent. This result suggests the need to consider other interpretations of the fricative perceptual spaces, and *auditory* interpretations is the most obvious of choices and easiest to implement. An additional advantage is that this may lead to a clarification of the correlations between different domains of speech processing.

Previous studies investigating the relationship between perceptual and auditory spaces on vowels have reported that the two spaces were highly correlated, regardless of the naturalness of the experimental materials examined (§II.2.1). However, in the case of fricatives the corresponding relationship is expected to be different, since the identity of fricatives is less easily discernible from their acoustic characteristics, than is the identity of vowels. It is hypothesised that the simpler the spectral patterns of stimulus sets, the higher the correlation between their perceptual and auditory spaces, thereby reflecting the simplicity and directness of the perceptual processing involved. As a corollary to the hypothesis, it has already been confirmed that the correlation between perceptual and auditory spaces for analogous nonspeech stimuli, consisting of only two spectral peaks

and zeroes, is extremely high (§III.4).

This study extends previous studies of the relationship between perceptual and auditory spaces on vowels by investigating not only the Euclidean auditory spaces but also the spaces based on the distance metrics applied to spectral gradient and the negative part of the second derivative of the spectrum. These metrics are reported to be more effective in emphasising the spectral parts which are of phonetic importance (§II.3.1). Also this study extends previous distance metric analyses of steady-state spectra to dynamic spectra, by using non-linear alignment (described in §III.3.2.2).

The auditory space of each stimulus set used in the perceptual tests, together with the results of canonical correlation analyses between perceptual and auditory spaces, are presented and discussed for each of the distance metrics in turn.

## 2 Euclidean distance metric

### 2.1 Auditory spaces and spectra

The auditory spaces are obtained in four main stages:

i) a simple 1/3-octave rate auditory transform is applied to the frequency axis of the acoustic spectrum. The 1/3-octave transform roughly simulates the non-linear spacing and bandwidth of the auditory filters. The resulting spectrum is smoothed by a bank of auditory filters. The filter bandwidths increase with the new 1/3-octave frequency scale, reflecting the loss of resolution as frequency increases.

ii) the intensity axis is transformed into a logarithmic scale, to reflect the non-linear loudness density pattern in the auditory periphery. The outcome is an auditory excitation pattern.

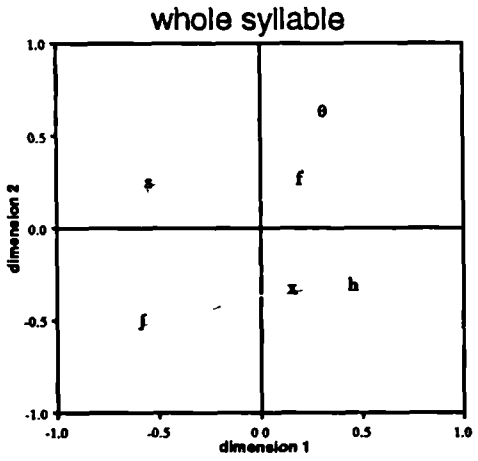
iii) spectral distances between these auditory excitation patterns are calculated using Euclidean metrics. A non-linear time alignment technique was used to account for the duration differences between the spectra.

iv) these distances are used as the input to 2-way MDS.

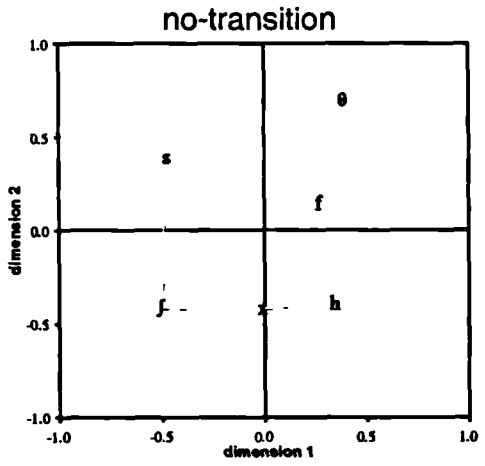
The same procedures will be followed for the other metric analyses, except the procedure shown above in (iv), in which the Euclidean metric will be replaced by other metrics. (For details of these procedures, see §III.3.2)

The auditory space for each stimulus set is presented in Figures V-1 (a) to (g).

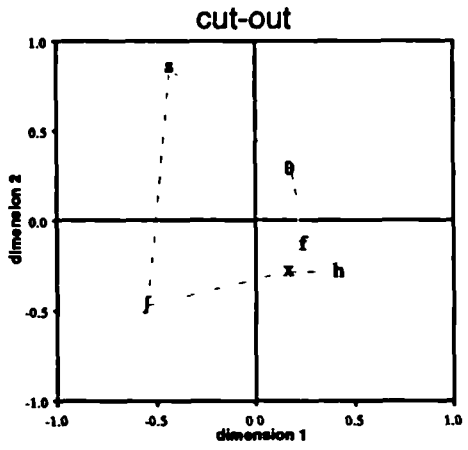
(a)



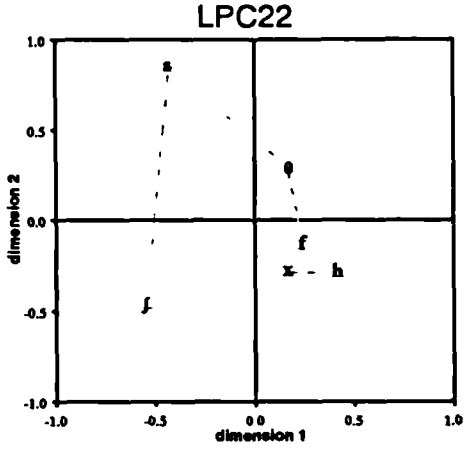

(b)



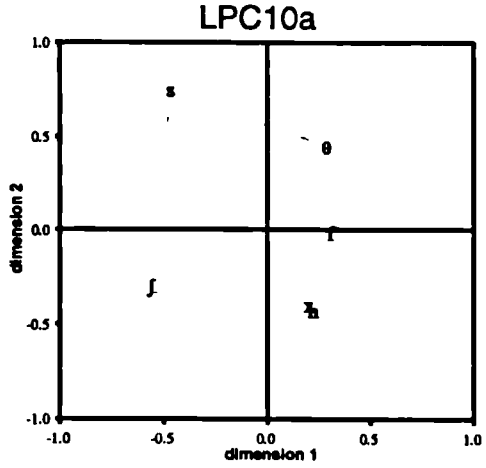

(c)



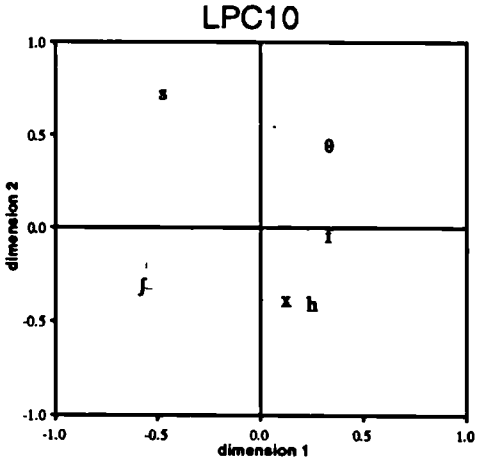

(d)



(e)




(f)




(g)

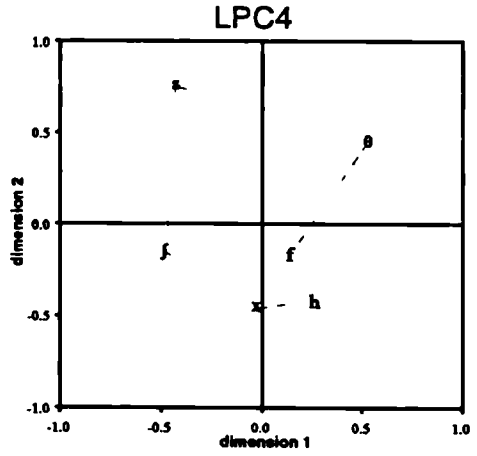



**Figure V-1.** Auditory spaces based on Euclidean distance metric. Fricatives are connected by broken lines in the order of /f θ s ʃ x h/.

The auditory configurations were based on the whole length of the stimulus, including the vowel sections in the whole syllable, no-transition, and LPC10a sets. Since the whole syllable set contained natural transitions and vowels, the auditory configurations based exclusively on the fricative portions were slightly different to those which included the whole syllable. However, the differences were very small. In the no-transition and LPC10a sets there were no transitions and exactly the same synthetic [ɑ] vowel was added

to each stimulus, thereby making an equal contribution to the spectral differences observed between the stimuli. Thus, vowel sections did not alter the auditory configurations in any significant way (compare Figures V-1 (e) and (f)).

The following patterns are noted:

- All the auditory configurations show a similar general pattern; /f θ/ occupy the upper right quadrant, /x h/ are in the lower right quadrant, and /s ʃ/ are located in the left half of the auditory spaces.
- Although the configurations change slightly from one set to the next, no clear trends are observable.

The cut-out and LPC22 sets are exceptions. In these sets, the fricatives /f/ and /x/ are positioned close to each other. As mentioned in §IV.2, the stimuli in the cut-out set are fricative portions of the stimuli in the no-transition set, with duration and intensity normalisation. The intensity normalisation — achieved by adjustment of overall RMS values — moves spectra up and down the amplitude scale, while preserving overall spectral shapes. The effects on the overall spectral shape of altering the duration with the program 'respeed' are not fully investigated before. However, the influence of duration adjustments on spectral shapes is investigated by comparing the average auditory spectra of each fricative in the no-transition and cut-out sets, presented in Figure V-2.

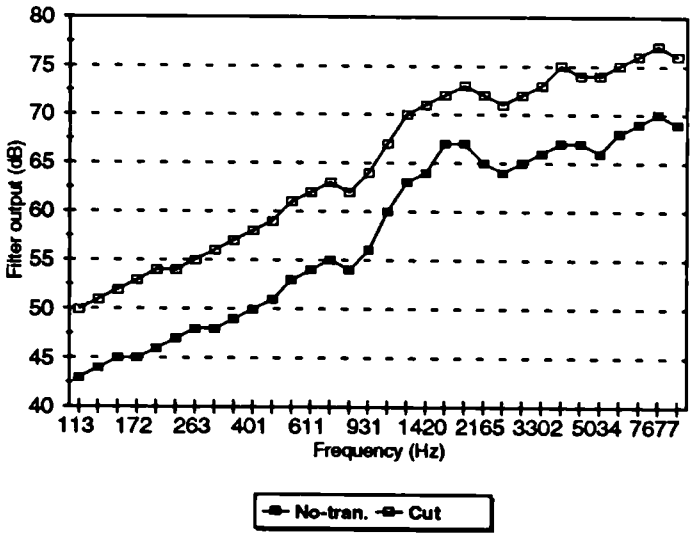
For /f/, the duration was only slightly adjusted, and therefore, the spectra of the cut-out and no-transition versions retain the same overall shape; they were raised by about 10 dB on the amplitude scale. For /θ/, the duration change has altered the spectral shape in high frequency regions. For /s/ and /ʃ/, the duration had to be shortened by about 30 ms, this having the effect of substantially altering the overall spectral shape. /x/ was the worst affected by the normalisation (shortened by 100 ms); the cut-out version (empty squares) shows rather flat spectral shape, while the no-transition version is rather peaky. Notice that the cut-out version of /x/ is very similar to that of /f/. This explains why the two fricatives are so close on the Euclidean map in Figure V-1 (c). /h/ had to be lengthened by 70 ms, but this did not affect its spectral shape.

Since the fricatives in LPC22 were modelled on the fricatives in cut-out set, the

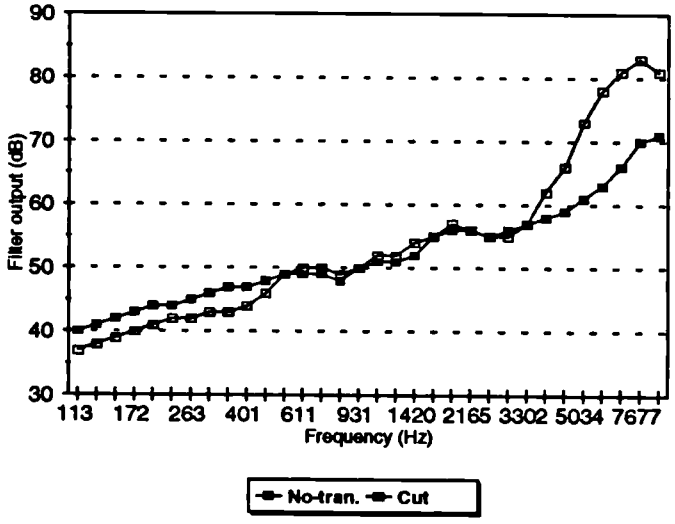
two respective Euclidean spaces are very similar. Although the sets LPC10 and LPC4 are also modelled on the same fricatives, the auditory spectra and Euclidean configurations are expected to be different from those for the cut-out set, since the number of peaks has been reduced.

Now that the spectral and spatial properties of fricatives in each set have been investigated, we shall next examine the correlation between the auditory and perceptual spaces.

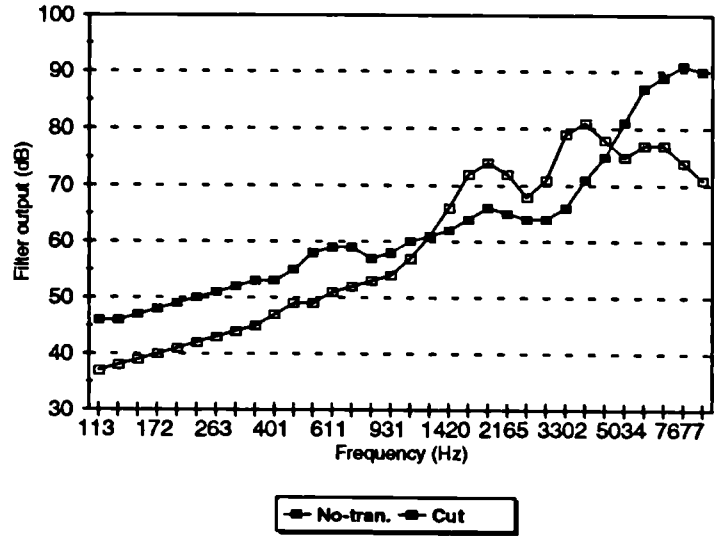
/f/



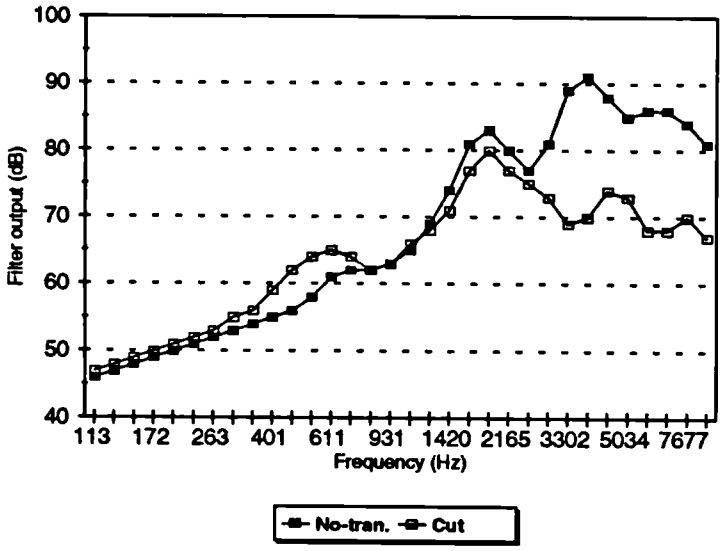

/θ/



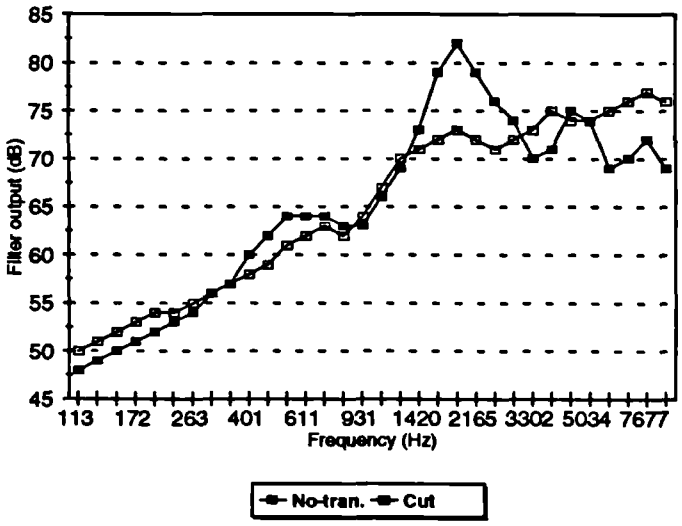
/s/


100
90
80
70
60
50
40
30
Filter output (dB)
113 172 263 401 611 931 1420 2165 3302 5034 7677
Frequency (Hz)
No-tran. Cut

/ʃ/


100
90
80
70
60
50
40
Filter output (dB)
113 172 263 401 611 931 1420 2165 3302 5034 7677
Frequency (Hz)
No-tran. Cut

*/x/*




*/h/*

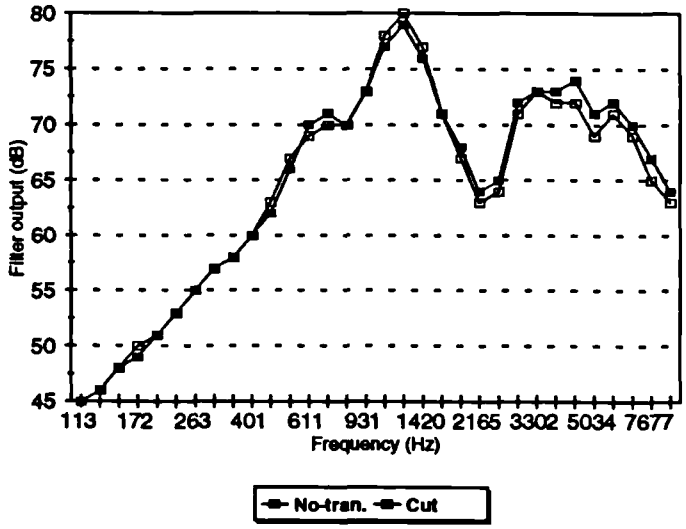



**Figure V-2.** 1/3-octave auditory spectra taken as an average over the whole length of the fricatives.

## 2.2 Canonical coefficients

Canonical coefficients between perceptual and auditory spaces are first reported as an indication of the perceptual/auditory relationship. Since the MDS analyses in the preceding chapter have shown that the similarity judgements of the two subject Groups A and B could be pooled together for the whole syllable, no-transition, and cut-out sets, the group perceptual spaces for all subjects were correlated with the auditory configurations. For the remaining stimuli sets, the perceptual configurations from the two groups were correlated separately with the auditory spaces. The auditory configurations were based on the spectral differences between the stimuli as a whole. A table summing the canonical correlations between the spaces and their significance is given below (Table V-1).

| Stimulus set | dimensions | Canonical coefficients | Significance |
|---|---|---|---|
| whole | 1 | 0.995 | 0.002 |
| syllable | 2 | 0.933 | 0.024 |
| no-tran. | 1 | 0.999 | 0 |
| | 2 | 0.908 | 0.037 |
| cut-out | 1 | 1.000 | 0 |
| | 2 | 0.744 | 0.156 |

| | | Group A | | Group B | |
|---|---|---|---|---|---|
| Stimulus sets | dimensions | Canonical coefficient | Significance | Canonical coefficient | Significance |
| LPC22 | 1 | 0.987 | 0.040 | 0.993 | 0.016 |
| | 2 | 0.653 | 0.232 | 0.746 | 0.148 |
| LPC10a | 1 | 0.974 | 0.027 | 0.992 | 0.005 |
| | 2 | 0.904 | 0.035 | 0.941 | 0.017 |
| LPC10 | 1 | 0.960 | 0.050 | 0.999 | 0.001 |
| | 2 | 0.878 | 0.050 | 0.945 | 0.015 |
| LPC4 | 1 | 0.966 | 0.034 | 0.990 | 0.005 |
| | 2 | 0.906 | 0.034 | 0.961 | 0.009 |

**Table V-1.** Canonical correlations between the perceptual and auditory spaces for each stimulus set.

- Overall, the correlations between the perceptual and auditory spaces are very high throughout, and are mostly significant[1] ($p < .05$). Exceptions are dimension 2 of the cut-out and LPC22 sets. Their significance values are also the lowest. A graphic representation of these correlations will be checked in §2.3 as additional justification for significance.

- There is no obvious trend in the correlations observed from one stimulus set to the next. The correlations in the whole-syllable set are almost as high as — or even higher than — those in the LPC4 set. This conflicts with the initial hypothesis (§1), that the correlations may be poorer for natural speech sounds than for synthetic ones, since the perception of speech sounds is clearly related to the phonetic properties of fricatives, and those phonetic properties cannot be described as a direct function of a few distinct acoustic properties. Since the auditory spaces of the different sets (in Figures V-1 (a) to (f)) were not related to phonetic properties, it is rather intriguing how the correlations between perceptual spaces were related to both phonetic and auditory spaces. A more detailed discussion of this relationship can only be carried out in terms of spatial representations. This will be undertaken in the next section.

- There are no clear differences in the results for the two subject Groups A and B, although the correlations are generally higher for Group B than for A. The improvement is, however, marginal. For example, the correlations for dimension 2 of the LPC22 set are visibly no better in Group B than in A, although the perceptual dimensions of Group A were phonetically interpretable while those of B were not. Also, these small differences in correlations may prove to carry no real significance, in view of the errors which may have been involved in the similarity judgements and nonlinear transformations of the judgments carried out in MDS. So, once again, correlations between perceptual and auditory spaces

[1] As it was already mentioned in §III.3.3, the 'real' significance may be slightly less than shown in Table V-1. This is because the MDS dimensions are not completely independent of each other, and thus, the assumption that the locations on the plane are independent random variables is not maintained, although it is likely that there is some linkage between the locations on the map.

were high, regardless of phonetic interpretability. That is, the perceptual organisations of Group B were not phonetically interpretable for the sets LPC22 and LPC10a, but this did not automatically give higher correlations with the auditory space. This contradicts the initial hypothesis.

In the next section, we shall examine the actual coordinate positions of the fricatives on the perceptual and auditory spaces, in order to ascertain i) the overall structure of the spatial relationship, and more specifically, ii) how some of the stimuli sets are related both to auditory and phonetic spaces.

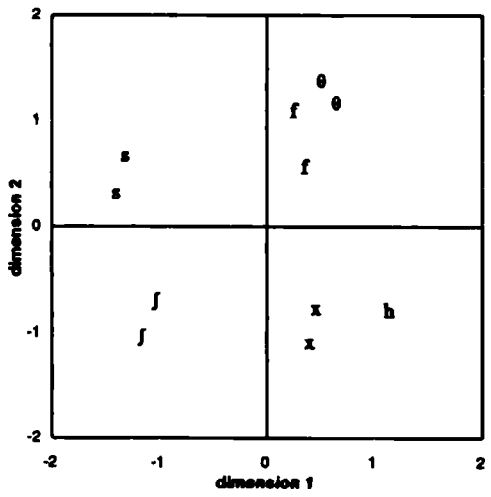
### 2.3 Canonical scores

Canonical scores provide a graphic representation of the relationship between the perceptual and auditory spaces, which had been scaled and rotated to give optimal correlations. The canonical scores for the first three sets are plotted in Figures V-3 (a) to (c).

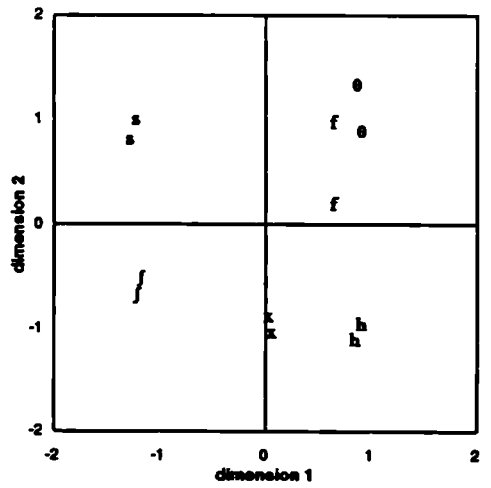
For the whole syllable set, the corresponding coordinates of the perceptual and auditory spaces are placed in a very close proximity; for /h/, they are, in fact, identical. The small differences observed between the two sets of coordinates are found mainly on the vertical dimension. These differences correspond to the second canonical correlation, which was found to be weaker than the first. A possible exception is /f/, which is further apart from /θ/ in the auditory configuration than in the perceptual one (refer back to Figure IV-2 (a)).

Overall, it is remarkable how closely related the auditory organisation of 'natural' fricatives is to their perceptual organisation. As it stands, this result implies that the relationship between auditory and perceptual spaces is the same for fricatives as for the vowel cases. However, the place dimension was not clear on the auditory space. For vowels, the F1/F2 space was closely related to the traditional phonetic vowel quadrilateral, which is loosely based on articulation. However, since the auditory space was based on the production of a particular speaker, the evidence is inconclusive. This suggests that a general auditory map of fricatives based on multiple speaker productions needs to be investigated. Also, the auditory modelling based on other distance metrics may

(a)




(b)




(c)

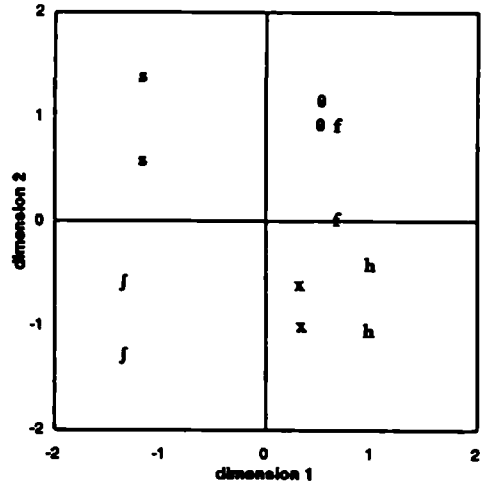



**Figure V-3.** Plots for canonical scores, comparing perceptual and Euclidean auditory spaces for (a) whole, (b) no-transition, and (c) cut-out sets.

give more accurate, and phonetically relevant results (in §3 and 4).

Exactly the same interpretation could be given for the no-transition set in Figure V-3 (b).

For the cut-out set, the coordinates of the fricatives are exactly the same on dimension 1 for both perceptual and auditory spaces (as indicated by the correlation coefficient, which was 1, in Table V-1). On dimension 2, the most noticeable feature is that the Euclidean distance between the fricatives /s/ and /ʃ/ is very large (referring back to Figure IV-10). Let us remind ourselves at this point that the spectral shape of /s-ʃ/ had changed a great deal after duration normalisation, comparing Figures V-2 (c) and (d). Overall, the perceptual and auditory spaces are very similar, despite the low correlation value on dimension 2, which was 0.744.

For the remaining four sets, Groups A and B are compared separately in Figures V-4 to 7.

For the LPC10a set, the perceptual data in Group A were phonetically interpretable, which was not the case in Group B. This interpretability depended upon the orientation of the axes on the perceptual map for Group A and the ordering of the fricatives on the 'place' dimension. However, when the configurations are rotated to find an optimal match with the auditory configurations as in Figures V-4 (a) and (b), these subtle phonetic differences could not be observed between the two groups. A noticeable feature of this set is that the Euclidean distances for /f/ and /h/ are rather inaccurate for modelling the perceptual distances, for both groups. Because the small number of stimuli (6), the mismatch in these two fricatives contributed significantly towards correlation values (.653 and .746 on dimension 2 for Groups A and B respectively).

For the sets LPC10a, LPC10, and LPC4, the matching between the perceptual and auditory configurations is rather impressive throughout, irrespective of whether or not a particular perceptual map was highly related to phonetic properties.

This rather conflicts with the hypothesis that the correlations between the perceptual and auditory maps based on the Euclidean metric will be higher for the stimuli sets of which the perceptual dimensions were not correlated with phonetic properties, thus reflecting the relatively direct perceptual processing involved in noise-like sounds. The Slope and N2D metrics are reported to be more sensitive to the phonetically relevant

properties of spectra, and thus, the auditory maps may show more phonetically orientated configurations. This is investigated in the following section.

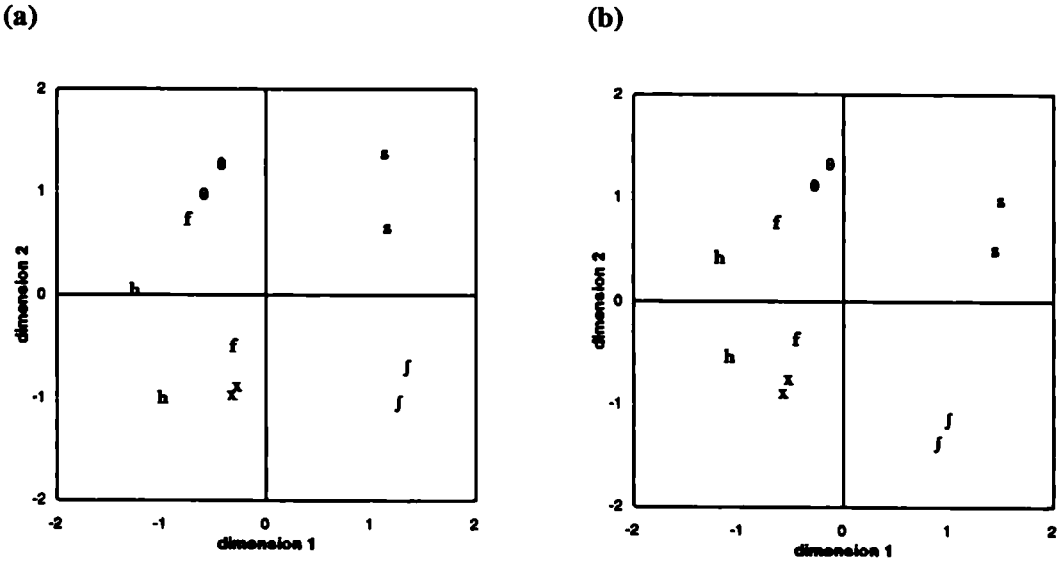



**Figure V-4.** Plots for canonical scores, comparing perceptual and Euclidean auditory spaces for (a) Group A, and (b) Group B, for the LPC22 set.

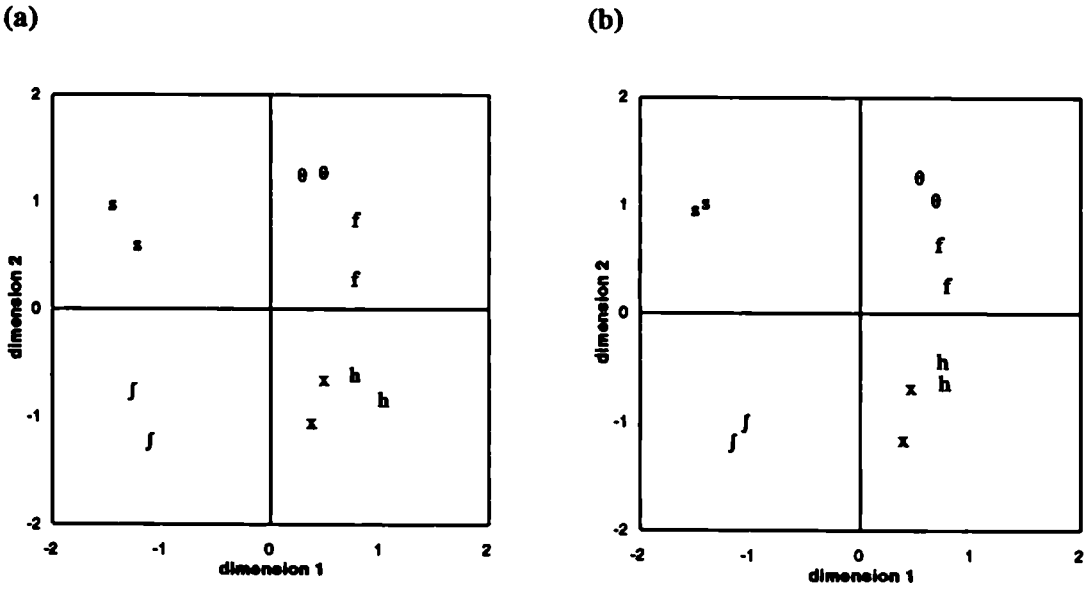



**Figure V-5.** Plots for canonical scores, comparing perceptual and Euclidean auditory spaces for (a) Group A, and (b) Group B, for the LPC10a set.

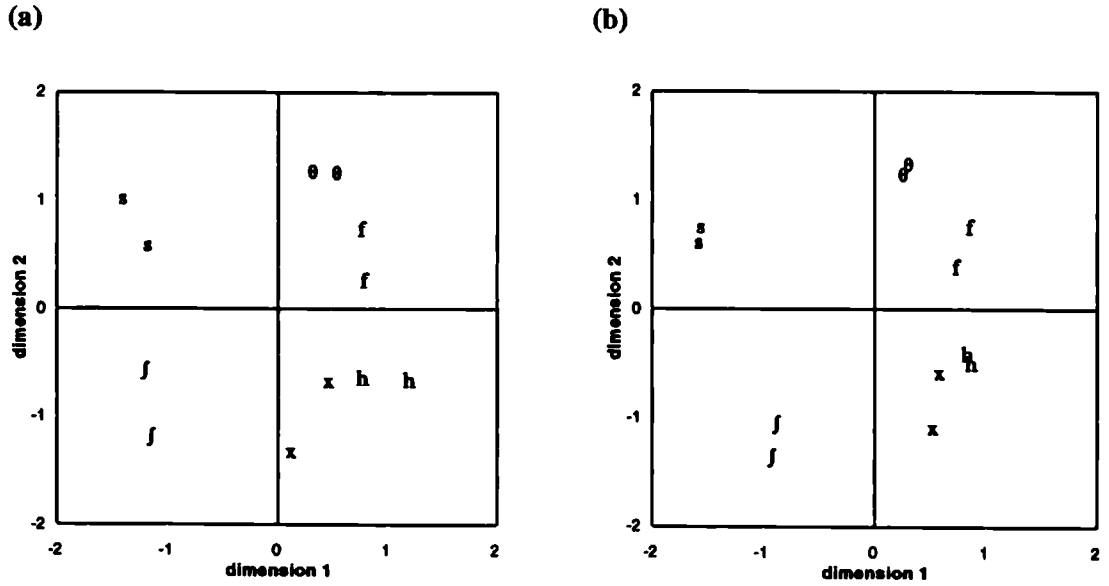



**Figure V-6.** Plots for canonical scores, comparing perceptual and Euclidean auditory spaces for (a) Group A, and (b) Group B, for the LPC10 set.

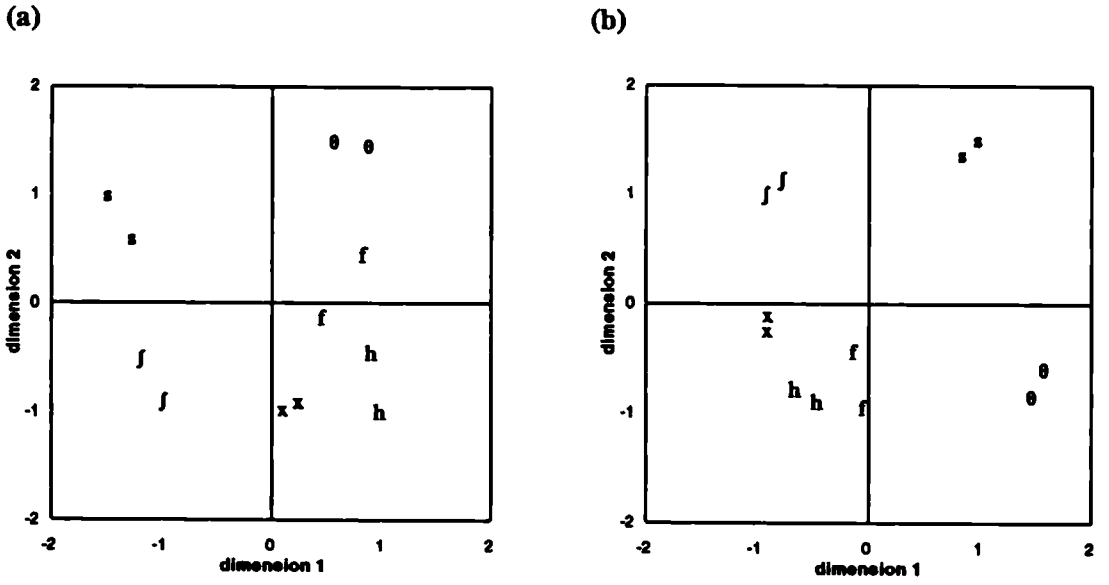



**Figure V-7.** Plots for canonical scores, comparing perceptual and Euclidean auditory spaces for (a) Group A, and (b) Group B, for the LPC4 set.

## 3 Slope distance metric

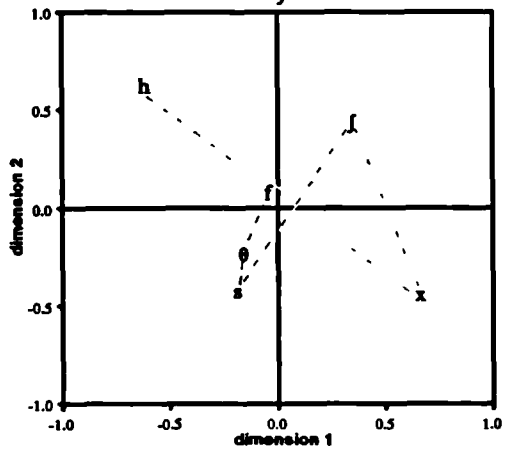
### 3.1 Auditory spaces

The procedures for obtaining the auditory spaces are the same as those described in §2.1, except for procedure (iv). For the Slope metric, Euclidean distances are calculated between the spectral gradients, rather than the level spectra. It was mentioned in §III.3.2.3 that the constants $k_E$, $k_{Gmax}$ and $k_{Lmax}$ in the Weighted Slope Metric need to be specified by the user, but this will only be carried out if the unoptimised form of the metric can yield meaningful results. Thus, for the moment only the unoptimised form of the metric is applied for the analyses. The resulting Slope configurations for the fricative stimuli are presented in Figures V-8 (a) to (g).

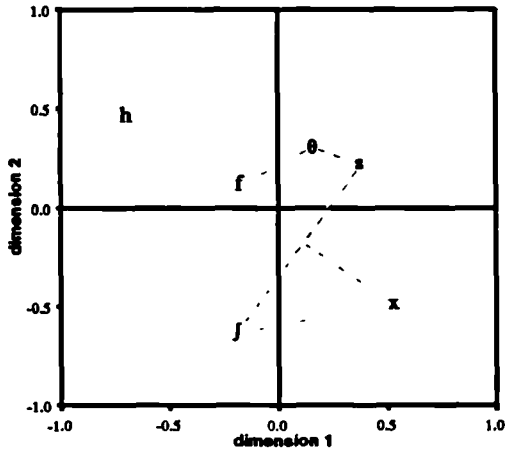
Not only is there no similarity between these and the Euclidean spaces we examined in Figure V-1, but also it is difficult to formulate any general description of the spaces. If we are pressed to identify any discernible pattern among the spaces, we could concede that /f / and /θ/ tend to stay close to each other. However, the lines joining the fricatives in the order /f θ s ʃ x h/ repeatedly intersect each other, showing random fluctuations in spectral distances from one fricative to another. The only exception is the LPC4 set (Figure V-8 (g)), where the configurations are comparable to those of the Euclidean space in Figure V-1 (g).

Because the unoptimised auditory spaces are completely uninterpretable, there seems little motivation for calculating the weighting functions of the metric.

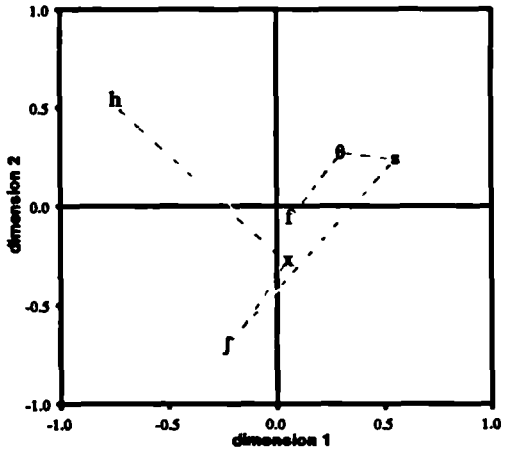
(a)


whole syllable
h
ʃ
f
θ
s
x
dimension 2
dimension 1
1.0
0.5
0.0
-0.5
-1.0

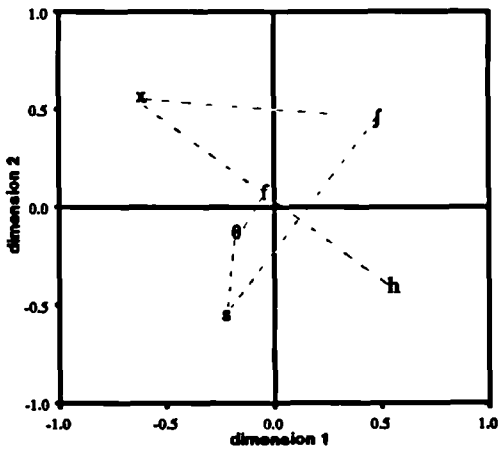
(b)


no-transition
h
θ
s
f
x
ʃ
dimension 2
dimension 1
1.0
0.5
0.0
-0.5
-1.0

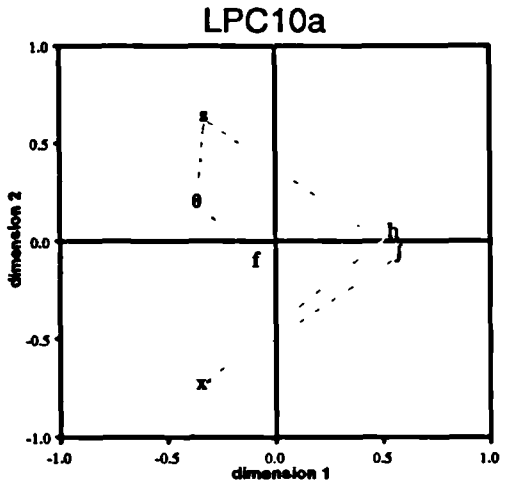
(c)


cut-out
h
θ
s
f
x
ʃ
dimension 2
dimension 1
1.0
0.5
0.0
-0.5
-1.0

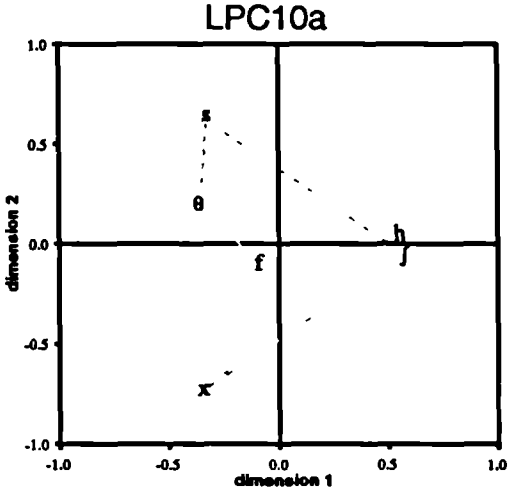
(d)


LPC22
x
ʃ
f
θ
h
s
dimension 2
dimension 1
1.0
0.5
0.0
-0.5
-1.0

(e)




(f)




(g)

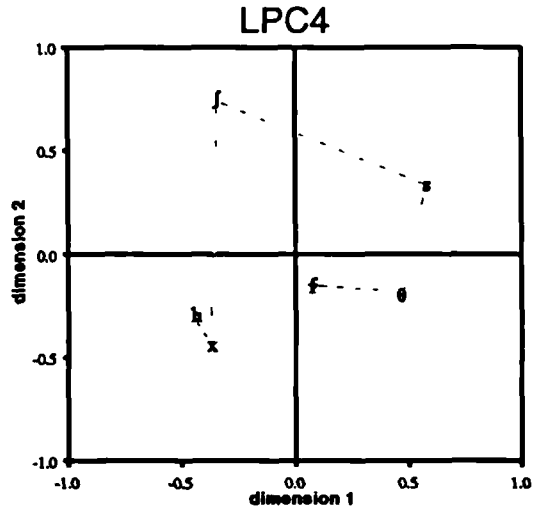



**Figure V-8.** Auditory spaces based on the Slope distance metric. Fricatives are connected by broken lines in the order of /f θ s ʃ x h/.

### 3.2 Canonical correlations

From the results of the auditory spatial analyses, it is predicted that the canonical correlations between the perceptual and Slope spaces will not be very high. However, canonical correlation analyses are carried out for the sake of completeness, and to show that the high canonical correlations observed, together with their significance in the Euclidean metric, are not the only possible outcome of such analyses. The results are summarised in Table V-2 below:

| Stimulus sets | dimensions | Canonical coefficient | Significance |
|---|---|---|---|
| whole | 1 | 0.596 | 0.831 |
| syllable | 2 | 0.377 | 0.536 |
| no-tran | 1 | 0.755 | 0.643 |
| | 2 | 0.385 | 0.527 |
| cut-out | 1 | 0.912 | 0.335 |
| | 2 | 0.321 | 0.598 |

| | | Group A | | Group B | |
|---|---|---|---|---|---|
| Stimulus sets | dimensions | Canonical coefficient | Significance | Canonical coefficient | Significance |
| LPC22 | 1 | 0.601 | 0.895 | 0.699 | 0.802 |
| | 2 | 0.005 | 0.993 | 0.041 | 0.948 |
| LPC10a | 1 | 0.860 | 0.465 | 0.971 | 0.116 |
| | 2 | 0.293 | 0.636 | 0.309 | 0.617 |
| LPC10 | 1 | 0.759 | 0.718 | 0.910 | 0.367 |
| | 2 | 0.057 | 0.927 | 0.129 | 0.836 |
| LPC4 | 1 | 0.860 | 0.412 | 0.959 | 0.077 |
| | 2 | 0.500 | 0.391 | 0.799 | 0.105 |

**Table V-2.** Canonical correlations between the perceptual and auditory spaces for each stimulus set, based on the Slope distance metric.

The correlations are generally low throughout, and show a noticeable degree of fluctuation between stimulus sets. The correlations are never significant.

Next, N2D metrics are examined.

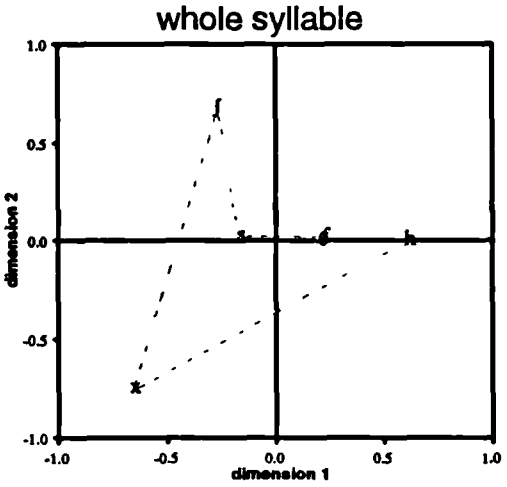
## 4 Negative second differential (N2D) metric

### 4.1 Auditory spaces

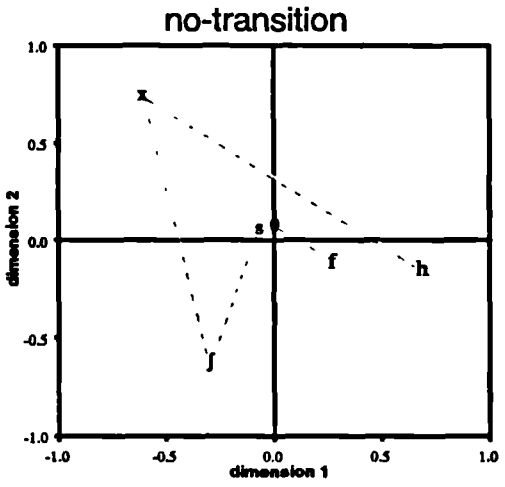
Procedure (iv) in §2.1 is replaced here by the distances based on the negative part of the second derivative of the spectra. As described in §III.3.2.3, the N2D metric thus explicitly eliminates spectral valleys, and instead emphasises peaks and shoulders of the spectra. As with the Slope metric, the unoptimised version of the metric is the first to be tried. That is, the Euclidean distances between the N2D of the spectra are used, without calculating the weighting functions. The resulting auditory spaces are presented in Figures V-9 (a) to (g).

As in the case of the Slope metric, the auditory spaces of the different stimulus sets are seen to fluctuate a great deal from one set to the next. For example, even between the sets LPC10a and LPC10 — where the only difference between the sets is the additional synthetic vowel parts in LPC10a — there are measurable changes in the overall distribution of the fricatives on the plane; for the set LPC10a, the fricatives /f θ ʃ/ are in the left half of the map, and the fricatives /s x h/ are in the right half of the plane; in contrast, for the LPC10 set the fricatives /f θ s/ are in the left half and /ʃ x h/ are in the right half of the plane. It may therefore be justifiably concluded that no meaningful interpretation of the N2D spaces is possible.

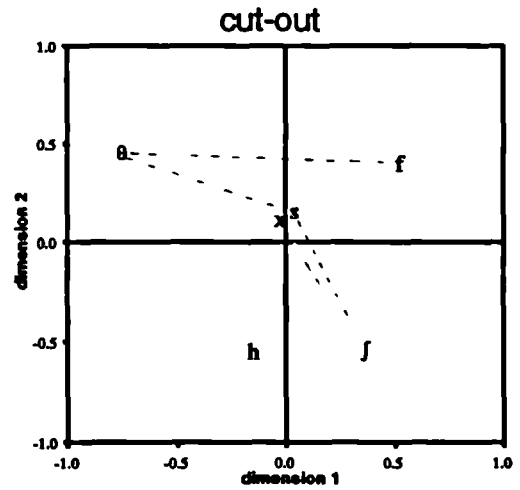
(a)


whole syllable
dimension 2
dimension 1
1.0
0.5
0.0
-0.5
-1.0
-1.0
-0.5
0.0
0.5
1.0

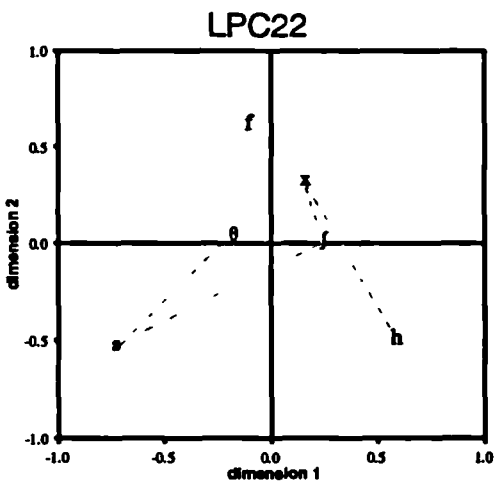
(b)


no-transition
dimension 2
dimension 1
1.0
0.5
0.0
-0.5
-1.0
-1.0
-0.5
0.0
0.5
1.0
s
f
h
ʃ

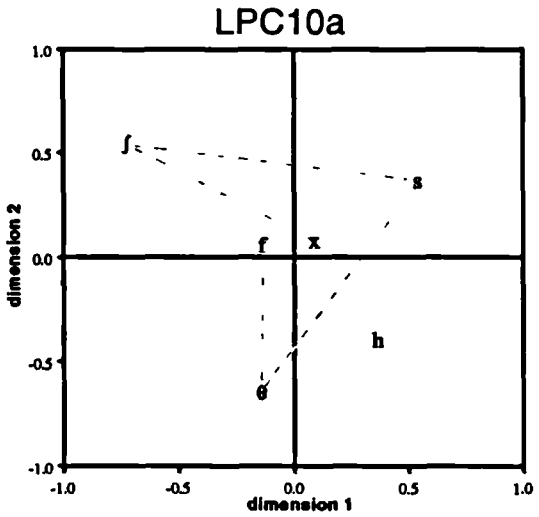
(c)


cut-out
dimension 2
dimension 1
1.0
0.5
0.0
-0.5
-1.0
-1.0
-0.5
0.0
0.5
1.0
θ
f
h
ʃ

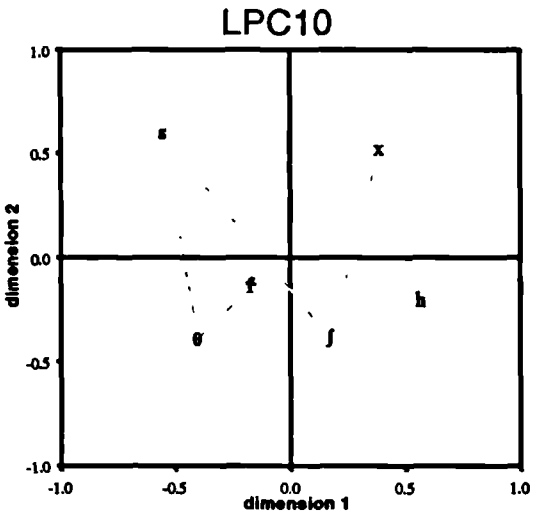
(d)


LPC22
dimension 2
dimension 1
1.0
0.5
0.0
-0.5
-1.0
-1.0
-0.5
0.0
0.5
1.0
f
θ
ʃ
h

(e)




(f)




(g)

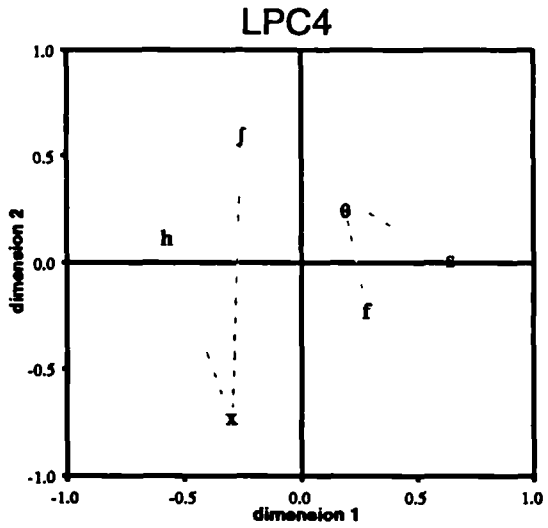



**Figure V-9.** Auditory spaces based on the N2D distance metric. Fricatives are connected by broken lines in the order /f θ s ʃ x h/.

## 4.2 Canonical correlations

By examining the auditory spaces we can predict that the canonical correlations will be neither high nor consistent. But in order to obtain a quantitative measurement of the relationship between the perceptual and N2D auditory spaces, the results of the correlation analyses are reported in Table V-3 below. From the results of the correlation analyses in Euclidean and Slope metrics, the subject groups A and B had given similar results. For this reason, only the results of the canonical correlation analyses on Group A are reported for the LPC synthesised stimuli.

| Stimulus sets | Dimensions | Canonical coefficient | Significance |
|---|---|---|---|
| whole | 1 | 0.935 | 0.279 |
| syllable | 2 | 0.175 | 0.779 |
| no-tran. | 1 | 0.954 | 0.156 |
| | 2 | 0.557 | 0.330 |
| cut-out | 1 | 0.958 | 0.178 |
| | 2 | 0.350 | 0.564 |
| | | Group A only | |
| LPC22 | 1 | 0.874 | 0.435 |
| | 2 | 0.347 | 0.567 |
| LPC10a | 1 | 1.000 | 0.001 |
| | 2 | 0.251 | 0.684 |
| LPC10 | 1 | 0.997 | 0.016 |
| | 2 | 0.365 | 0.545 |
| LPC4 | 1 | 0.943 | 0.250 |
| | 2 | 0.181 | 0.771 |

**Table V-3.** Canonical correlations between the perceptual and auditory spaces for each stimulus set, based on the N2D distance metric.

The results are self-explanatory, or rather, they lack any explanatory value; they also conform to the predictions made above, since the values of the canonical coefficients fluctuate and are generally of little significance (except for the first correlations of the

stimulus sets LPC10a and LPC10).

## 5 Discussion

### 5.1 Distance metrics

A comparison of the three different metrics shows the Euclidean metric to be the most accurate at predicting the perceptual results. That is, the canonical correlations between the perceptual and auditory spaces were highest when the Euclidean metric was implemented. This fails to support the results obtained for the vowel sounds reviewed in §II.3.1, where it was reported that the metrics which emphasize peaks and shoulders of the spectra predict the perceptual data more successfully.

Two possible reasons for this discrepancy may be put forward: a) the metrics are only effective for the smooth vowel spectra, and not for the fricatives, in which the speech signal fluctuates rapidly; b) the previous results were only true for the specific sets of stimuli used for their experiments.

In order to investigate these two possible reasons, the metrics were applied to a set of English vowels /i: e ɑ: ɔ: u:/, pronounced in the context /h_d/. A male speaker's production was taken from a large data base. Procedures for auditory analyses were the same as in §2.1. The auditory maps based on Euclidean, Slope, and N2D metrics are presented in Figures V-10 (a) to (c).

Figure V-10 (a) shows the Euclidean space of the five vowels. The configuration is comparable to the Euclidean space of 11 Dutch vowels in Pols *et al.* (1969). However, as in the case of the fricatives, the configurations from the Slope and N2D metrics do not follow the expected pattern of compatibility with the F1/F2 space. An increase in the number of filter channels to 64 and 128 did not improve the configurations of the Slope and N2D metrics.

Figure V-10' (a) lists the frequencies of the first and second formants of the English vowels used. The formants were determined from LPC modelled spectra. The formant values were plotted on a logarithmic scale, as shown in Figure V-10' (b). The formant configurations closely resemble the Euclidean map in Figure V-10 (a). The canonical correlations, in Figure V-10' (c), supports this observation.

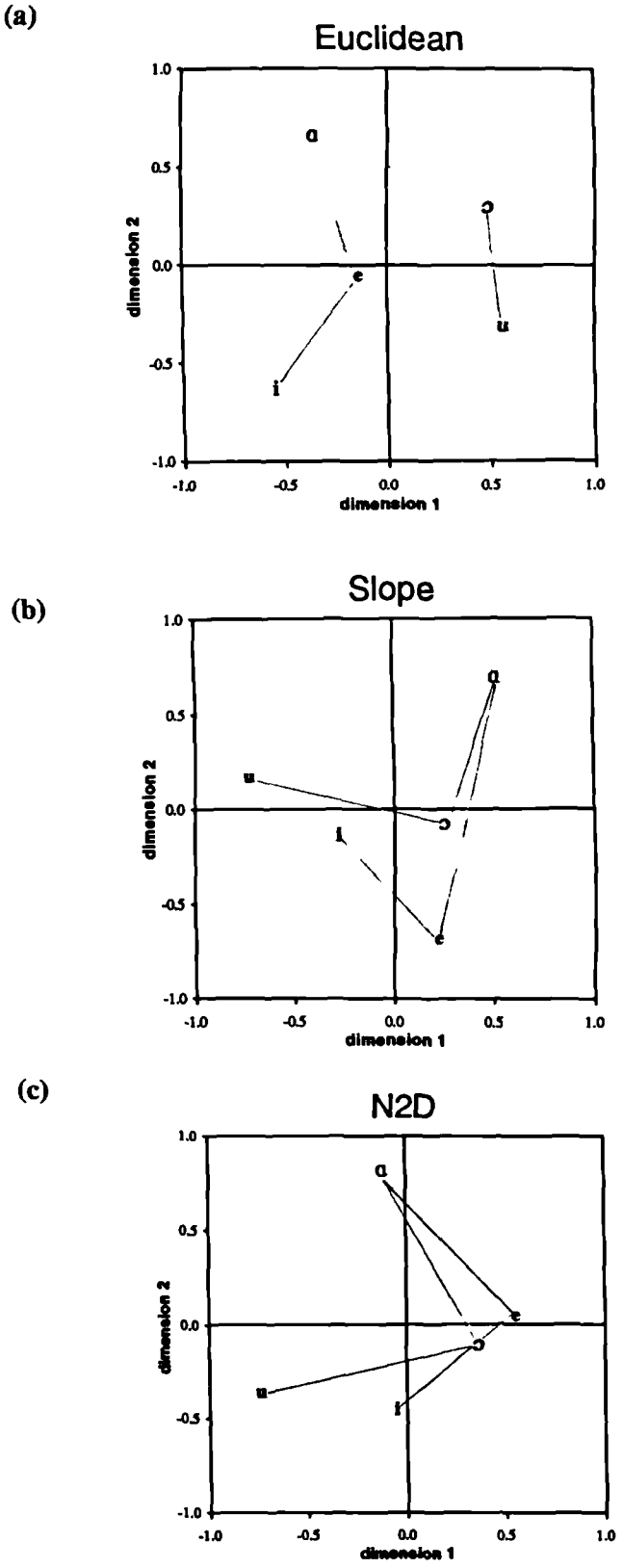



**Figure V-10.** Five vowel auditory spaces from (a) Euclidean, (b) Slope, and (c) N2D metrics..

Therefore, the results show that the metrics are also ineffective in representing the auditory space of the vowels themselves, pointing to the reason given here as (b) for the discrepancy with previous studies that the metrics are only useful for the specific sets of stimuli used in their experiments, which may have been rather artificial.

Indeed, Klatt's 1982(a) study used 66 variations of the vowel /a/ each differing subtly in terms of acoustic properties. Klatt found that pairs of synthesised /a/ vowels which differed in terms of their formant frequencies were given the highest distance scores in a '10-point scale' of 'phonetic distance' judgements (as opposed to the 'psychoacoustic distance'). This means that he needed to devise a distance metric which would emphasize the formant frequencies, whilst ignoring other spectral variations. However, when the data involve different vowels with clearly different formant positions in the spectra, it is possible that the metric may be over-emphasising the differences. Therefore, the metric may be rather specific to the particular stimulus type used[2].

Also in the Assmann & Summerfield study (1989), there was concern over how well the pattern-matching procedure based on different distance metrics predicted the vowel identifications in the presence of competing voice (simultaneous double vowels). This means that they may have also needed to give extra emphasis to the spectral peaks, in order to allow each vowel to stand out from the other in the double vowels.

Besides, Bakkumm *et al.* (1993) report that the Euclidean metric with spectral weight optimised by each filter band weight for each vowel category offered the highest correlation with the perceptual data. The perceptual data were collected by triadic comparisons of the pronunciation quality of 15 Dutch monophthongs and diphthongs, pronounced by native, foreign, and deaf speakers. Klatt's Slope metric performed badly in comparison with the Euclidean metric. In the light of this result, Bakkumm *et al.* concluded that "Obviously, the values of the various distance measures and weighting functions depend strongly on the specific tasks and analysis methods." (p1998).

Furthermore, the outputs of the Slope and N2D metrics have not been transformed to MDS dimensions in previous studies, thus, the results cannot be fully compared.

The adequacy of Euclidean distance modelling will be further investigated with

---

[2]Nocerino *et al.* (1985) obtained highest recognition rates for the unoptimised Slope metric. However, in recognition tasks, parameters besides the metrics themselves are seen to interact.

multiple speaker data in the next chapter.

## 5.2 The relationship between perceptual and auditory spaces

The relationship between perceptual and auditory spaces is discussed on the basis of the results of Euclidean distance metric modelling. The results showed that the canonical correlations between them were high for most of the stimulus sets, regardless of the degree of naturalness in the stimuli. This was rather at odds with the hypothesis that the perceptual map of the whole syllable set which manifested phonetically interpretable dimensions would be less highly correlated with the corresponding auditory map than the perceptual map of LPC4 or LPC10.

However, the auditory maps never clearly manifested phonetically interpretable dimensions for this particular speaker's production of fricatives. There are two possible interpretations for these results:

a) the MDS technique is not sensitive enough for showing subtle changes in the relationship between perceptual and auditory spaces from natural to synthetic fricatives or;

b) the *general* auditory map (based on productions of multiple speakers) of fricatives may be closely related to phonetic properties; thus, as in vowels, there is simple correspondence between phonetic, perceptual and auditory spaces.

It is largely point (b) above which motivates the production tests presented in the next chapter.

# *Chapter VI. Production tests*

## 1 Introduction and objectives

The objective in this chapter is to establish a general auditory space for English fricatives. To achieve this goal, we have carried out a background study of *production* in which intra- and inter-speaker variations are investigated. This study of speaker-dependent variation adopts two distinct approaches: i) variations *within* a single category of fricative were investigated in terms of auditory *spectral* representations, leading to an average auditory spectral shape for each fricative type (in §3); ii) variations *across* fricative categories are investigated in terms of *spatial* representations of all productions, leading to a general auditory map of fricatives (in §4). An average auditory map of all productions is also shown, and the spatial properties of the stimuli used in the perceptual tests (Chapter IV) are compared with the results of these production tests. Finally, the acoustic correlate to each auditory dimension is identified.

A possible explanation for the unexpected results in Chapter V may be found if the materials used in the perception tests were atypical of English fricatives, or if the general auditory space of fricatives display dimensions which clearly correlate with phonetic properties.

## 2 Materials: recordings and speakers

Five male native speakers of English in the 20-40 age group recorded the fricatives, /f θ s ʃ h/, followed by the vowel [ɑ]. They were asked to utter the syllables twice, clearly and in a falling tone. The recordings were made in an anechoic room onto a Sony DTC-1000ES digital audio tape recorder. They were digitised with a 16-bit quantization rate and a sampling rate of 20 kHz.

## 3 Fricative spectra

### 3.0 Introduction

In the initial stage of the production study, various spectral properties were measured and

summarised. Although the ultimate objective is to obtain a general auditory space, an examination of the basic spectral properties of fricatives, and how these may bear on spatial representations, is carried out as a background study. These spectral properties may help us to understand the spatial variations among different speakers and productions.

### 3.1 Initial measurements (duration and intensity)

From each of the recorded syllables the fricative portions were marked out to exclude the transition section. The method for this process is described above in §IV.2.

The duration of the marked fricative sections was measured by a simple computer program. The means of the duration measurements for each fricative type are given in Table VI-1.

| fricatives | f | θ | s | ʃ | h |
|---|---|---|---|---|---|
| mean | 138.8 | 142.3 | 202.7 | 189.6 | 115 |
| s.d. | 37.01 | 39.89 | 32.23 | 30.76 | 38.84 |

**Table VI-1.** The mean duration, in ms, of the fricative portions of two repetitions from the five different speakers.

Previous studies have reported that the duration of fricatives increases progressively in the following order: dentals, labials, alveolars, and palatals (You, 1979). Similar pattern was also observed here (except that /θ/ was longer than /f/, and /s/ was longer than /ʃ/ — but since the standard deviations are larger than the differences in the length, these differences are not significant). Traditionally, the durational differences between the fricatives have not been considered to bear any direct linguistic significance. Thus, this property is not analysed further.

RMS (root mean square) values of the whole length of the fricative sections were measured as an indication of amplitude levels. The whole length of the fricatives was taken, since fricatives have random energy fluctuation, so that no single part of any segment could be extracted and used to represent the intensity. Two-way ANOVA was carried out in order to investigate the subject and fricative variations across all the productions of different fricatives. The null hypothesis was that the mean values are

identical across all subjects and fricative types. The obtained F-ratio for subjects was $F(4,25) = 1.16$, $p \geq .351$; it cannot reject the null hypothesis, which implies that there is no subject effect. On the other hand, the obtained F-ratio for different fricative types indicates that there is significant fricative type effect; $F(4,25) = 5.82$, $p \leq .002$. Thus, RMS levels of each fricative type can be averaged across all the speakers, as presented in Table VI-2.

| fricatives | f | θ | s | ʃ | h |
|---|---|---|---|---|---|
| mean | -41.87 | -39.92 | -25.37 | -25.28 | -35.39 |
| s.d. | 2.89 | 2.61 | 4.40 | 3.02 | 4.16 |

**Table VI-2.** The mean RMS levels of fricatives across different productions, measured from the whole length of the fricative sections.

Since the RMS level is calculated in relation to the RMS level of a sine wave that fills the 16 bit range, the RMS values here are negative. As expected, the mean RMS levels of /s ʃ/ are much higher than those of the other fricatives, at about -25 dB; /f θ/ are about 15 dB lower; and the average RMS value of /h/ is about 10 dB lower than /s ʃ/.

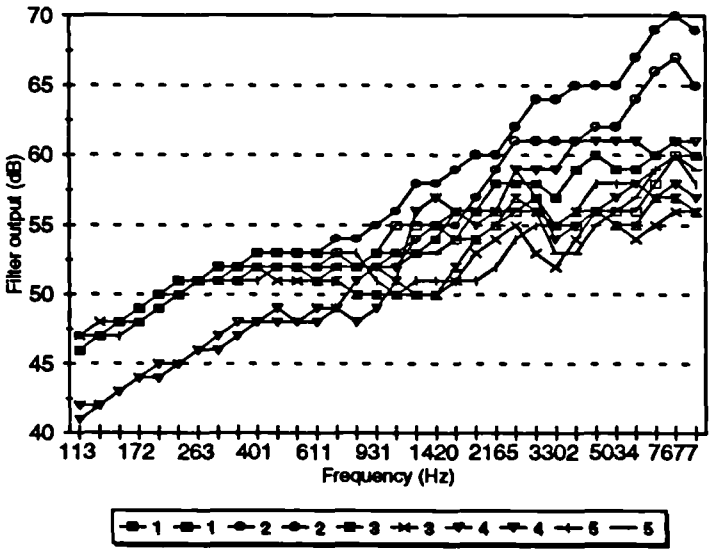
Each RMS level measurement was later used for loudness normalisation. The effect of different loudness levels on spectral and spatial variations among speakers is examined in the following sections.

### 3.2 Auditory spectra

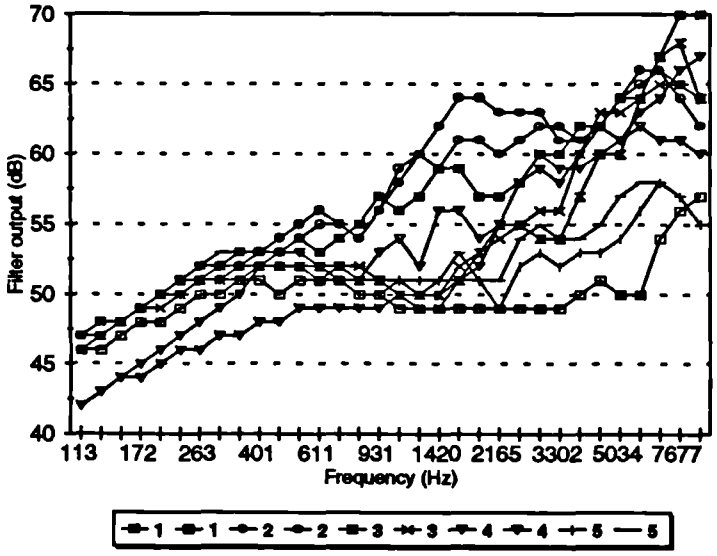
Each fricative segment was analysed by 1/3 octave bandpass filtering. There were 32 filters, ranging from 100 to 9000 Hz. The output energy levels of each filter was averaged across the whole length of each fricative segment. In this way, for every individual production, a series of 32 numbers was obtained, representing 32 filter bands.

Figures VI-1 (a) to (e) show the time-averaged auditory spectra of each fricative type, plotted on the same axes to assess the speaker/production-dependent variations. The horizontal axis represents the centre frequencies of the 32 filters in Hz. The vertical axis represents the energy levels of each filter in dB.

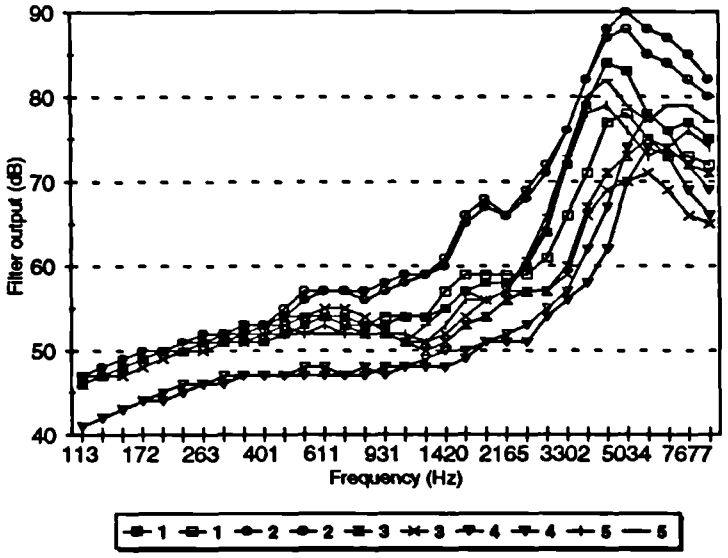
/f/


70
65
60
55
50
45
40
Filter output (dB)
113 172 263 401 611 931 1420 2165 3302 5034 7677
Frequency (Hz)
1 1 2 2 3 3 4 4 5 5

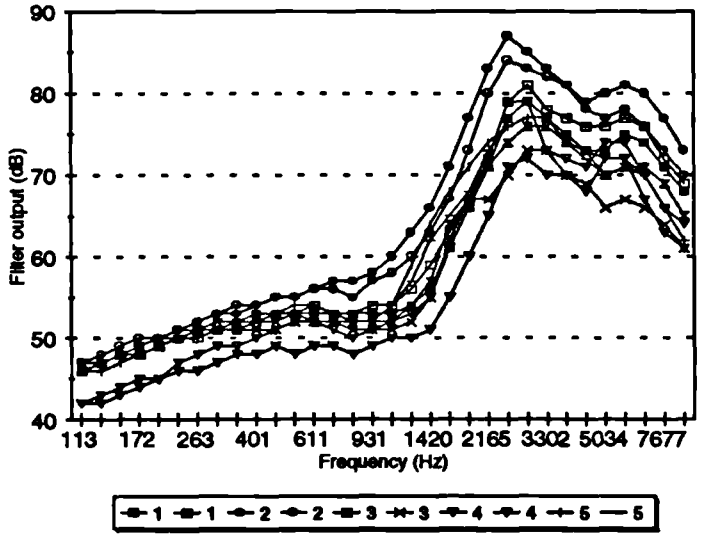
/θ/


70
65
60
55
50
45
40
Filter output (dB)
113 172 263 401 611 931 1420 2165 3302 5034 7677
Frequency (Hz)
1 1 2 2 3 3 4 4 5 5

*/s/*


90
80
70
60
50
40
Filter output (dB)
113 172 263 401 611 931 1420 2165 3302 5034 7677
Frequency (Hz)
1 1 2 2 3 3 4 4 5 5

*/ʃ/*


90
80
70
60
50
40
Filter output (dB)
113 172 263 401 611 931 1420 2165 3302 5034 7677
Frequency (Hz)
1 1 2 2 3 3 4 4 5 5

/h/

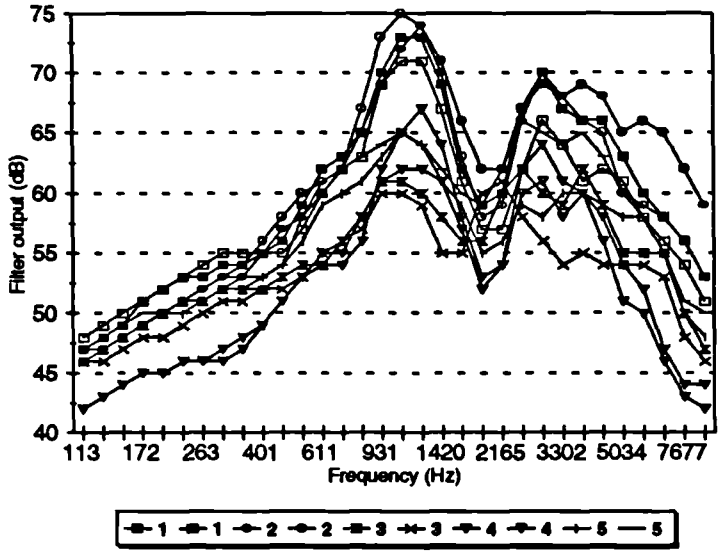



**Figure VI-1.** 1/3 octave auditory spectra of 10 productions by 5 speakers, taken as an average over the whole length of the fricatives.

Although it is difficult to separate one production curve from another, it is clearly observable, in Figure VI-1, that some of the spectral curves of /f/ have prominent peaks in the high frequency region (7000 Hz) — notably for speaker 2 (curves with circles), while other curves show a flat spectral shape in the same region — for example, for speaker 3 (curves with crosses). This kind of variation is also seen in the spectra of /θ/. For instance, in filter bands 20 to 25 (1600-3300 Hz), the spectra of speaker 2 show high energy output, while the second production of speaker 1 (curves with empty squares) shows a completely flat spectrum.

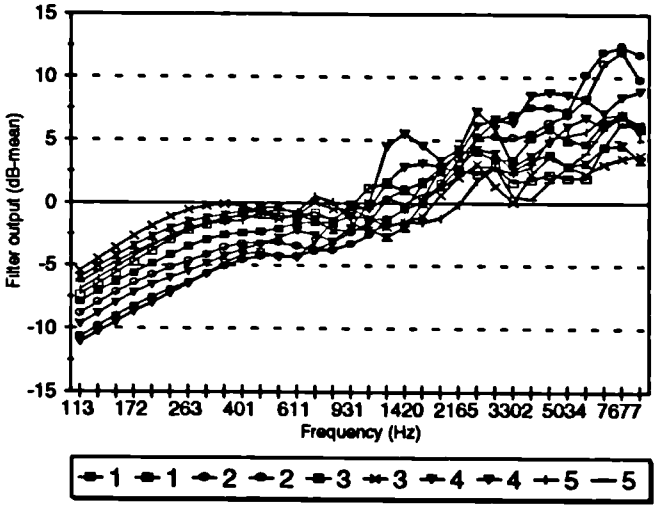
The 1/3 octave bandpass analysis curves for /s/, however, are generally of a similar shape, although there are variations in the degree of peakiness; the peak energy difference between speaker 2 and speaker 3 is about 20 dB. The same is true with the curves for /ʃ/, except that even less variation in spectral shape is discernible among them. The curves for /h/ also have same overall outline, mostly with two prominent spectral peaks.

From these observations, it can be speculated that the major source of variation among the spectra of the different productions of each fricative may be the overall energy level differences.

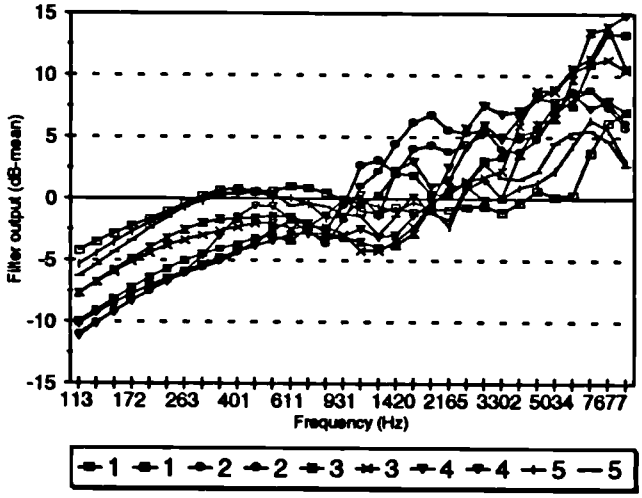
In order to accommodate the differences in the overall level of the fricative segments, the output levels of the 32 bands were reduced by the mean level of that particular production. This process was repeated for each production of each fricative.

Figures VI-2 presents the spectral curves of 1/3 octave analyses of fricatives, after subtracting the mean spectral energy in dB.

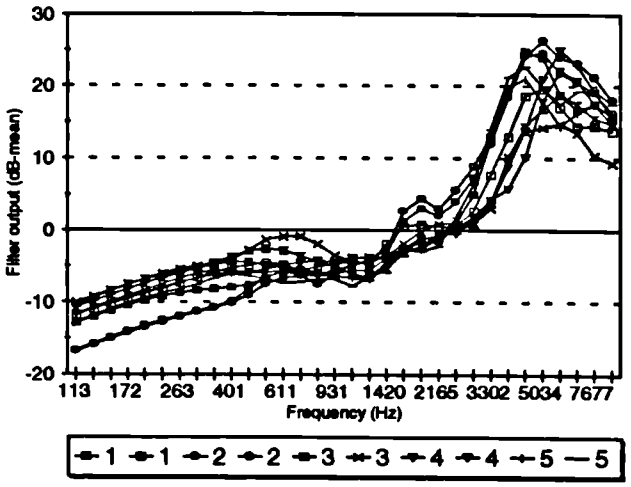
/f/



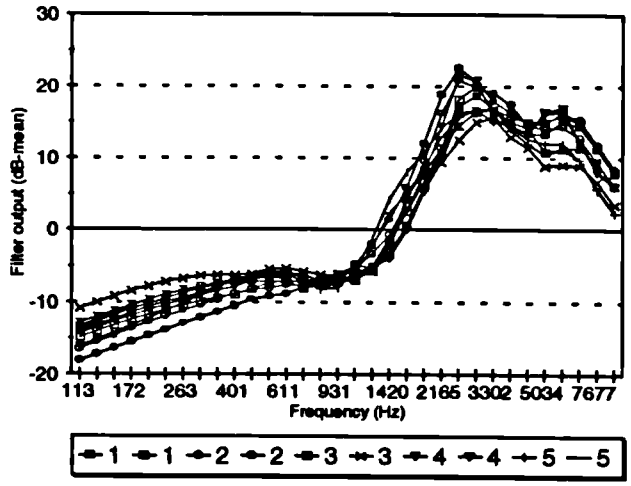

/θ/



/s/




/ʃ/




/h/

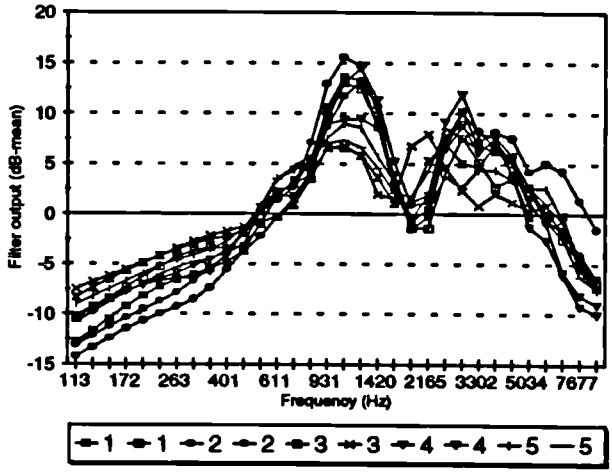



**Figure VI-2.** 1/3 octave auditory spectra of 10 productions by 5 speakers, taken as an average over the whole length of the fricatives, after subtracting the mean in dB.

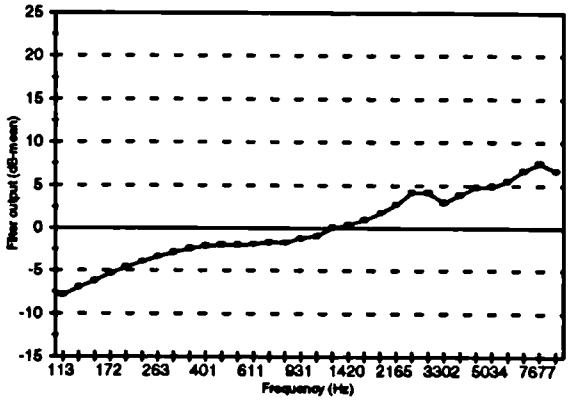
These graphs show that, after normalisation, there is very little production/speaker-dependent variation in the spectra of each fricative. For /f/ and /θ/ in Figure VI-2, it is now apparent that there are no prominent spectral high energy regions; instead, what we find is a tendency to increase energy output as the filter band frequencies increase. Variation arises as a result of a number of small spectral humps observed in the high frequency regions for different productions.

For the fricatives /s/ and /ʃ/, the general shape of the fricative spectra is very stable. This seems also true for /h/, although there are some variations in the exact locations of spectral peaks in the different productions.

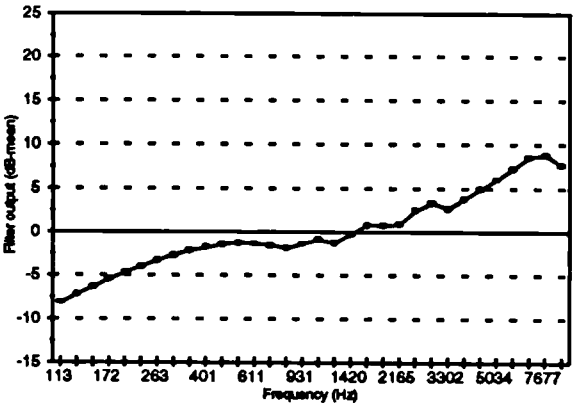
In sum, there is a high level of agreement among the different spectral curves of any one fricative.

Figure VI-3 represents the 1/3-oct frequency spectra of the fricatives, averaged over the ten productions spoken by the five different speakers. These are compared with the spectra of stimuli used in the perception tests. It is noticeable that the spectral characteristics of /f/ and /θ/ are very similar; in both cases, the spectra are mainly flat, but two small peaks are observed. For both fricatives, the first peak occurs between 1800 and 2000 Hz and the second peak occurs at around 5500 Hz. This is quite different from the spectra in Figure V-2, in which we observed that /f/ had a much flatter spectral shape than /θ/, and therefore was rather different from /θ/. /s/ and /ʃ/ can be characterised by a single broad-band peak; however, the low cut-off frequency occurs a little higher for /s/ at around 3600 Hz, than for /ʃ/, at 2000 Hz. For /h/, the spectral peaks occur at around 770 Hz and 2000 Hz, which correspond to the formant frequencies of the following [ɑ] vowel. These characteristics were also observed in Figure V-2.

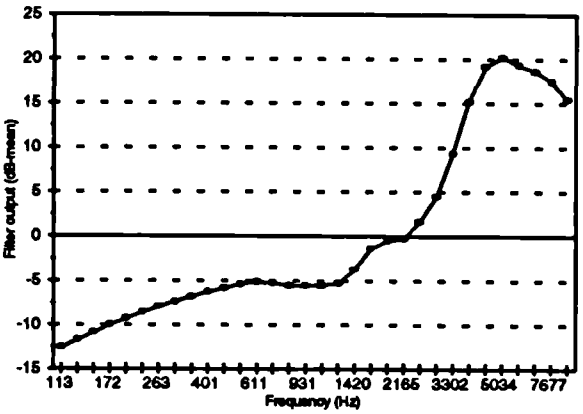
*/f/*


25
20
15
10
5
0
-5
-10
-15
Filter output (dB-mean)
113 172 263 401 611 931 1420 2165 3302 5034 7677
Frequency (Hz)

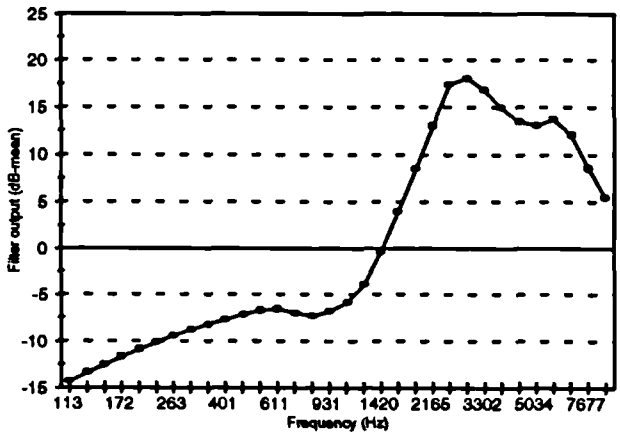
*/θ/*


25
20
15
10
5
0
-5
-10
-15
Filter output (dB-mean)
113 172 263 401 611 931 1420 2165 3302 5034 7677
Frequency (Hz)

*/s/*


25
20
15
10
5
0
-5
-10
-15
Filter output (dB-mean)
113 172 263 401 611 931 1420 2165 3302 5034 7677
Frequency (Hz)

/ʃ/




/h/

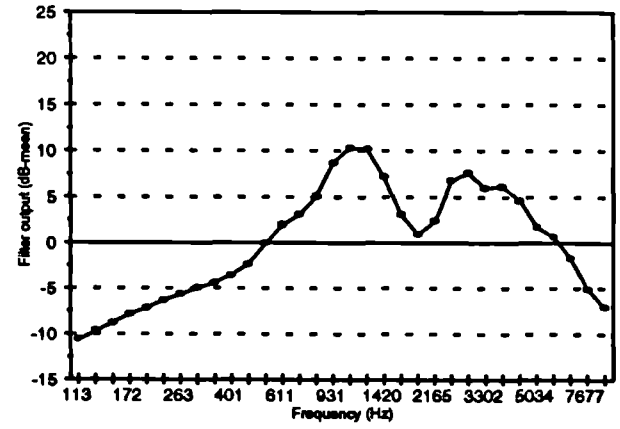



**Figure VI-3.** 1/3 octave auditory spectra, taken as an average over 10 productions by 5 speakers.

## 3.3 Summary

The initial measurements of fricative duration and loudness showed that, for both of these acoustic properties, the sibilant vs. nonsibilant distinction could be clearly identified. Speaker-dependent variation was not significant (§3.1).

There was some degree of spectral variation between different productions of the same fricative. The overall intensity normalisation showed that this variation was largely due to intensity differences rather than to any actual variation in spectral shapes. Again, inter- and intra- speaker variations were relatively small.

An informal comparison with the average auditory spectra of the stimuli used in the main perception tests (Figure V-2), we find that /f/ and /θ/ in perceptual tests were

much more distinct from each other than in the general auditory spectra (Figure VI-3). The spectra of the other fricatives are similar.

In the next section, the variation across fricative types is examined by relative spatial configurations of these fricatives on a two-dimensional plane.

## 4 Fricative spaces

### 4.0 Introduction

From the spectral analysis of the fricatives, it was observed that, after the spectral intensity normalisation, the spectral variation among the multiple productions by different speakers was very small. In this section, we shall examine whether similar intra- and inter-speaker variations are observed in the spatial relationships across the fricatives, in order to establish a general auditory map of fricatives. The procedures employed to obtain the auditory space for each speaker's productions were the same as those described in §V.2.1.

### 4.1 Intra-speaker variations

The metric 2-way MDS analysis was applied to each fricative set of each speaker independently. The mean accounted variances was 0.91 for dimension 1, and 0.08 for dimension 2. Thus, the two-dimensional solution was adequate to account for nearly all the variance. The auditory coordinates of the different productions of each speaker are used for canonical correlation analyses between the two separate productions of the same speaker. The canonical coefficients are presented in Table VI-3. The configurations of the two separate productions of the same speaker were rotated for optimal congruence, and the new sets of coordinates (canonical scores) are presented in Figures VI-4 (a) to (e).

| Speakers | Dimensions | Correlation | Significance |
|---|---|---|---|
| 1 | 1 | 0.971 | 0.075 |
| | 2 | 0.969 | 0.040 |
| 2 | 1 | 0.999 | 0.006 |
| | 2 | 0.989 | 0.017 |
| 3 | 1 | 0.987 | 0.111 |
| | 2 | 0.856 | 0.160 |
| 4 | 1 | 0.983 | 0.070 |
| | 2 | 0.952 | 0.059 |
| 5 | 1 | 0.998 | 0.029 |
| | 2 | 0.893 | 0.122 |

**Table VI-3.** Canonical correlation coefficients between the two different productions of each speaker.

The canonical correlation values show a reasonably close agreement between the two different productions of each speaker, although only speaker 2 shows a truly stable pattern between the two separate productions (See Figure VI-4 (b)). The lowest correlation was observed for speaker 3; the variations become more apparent in the spatial configurations shown in Figure VI-4 (c).

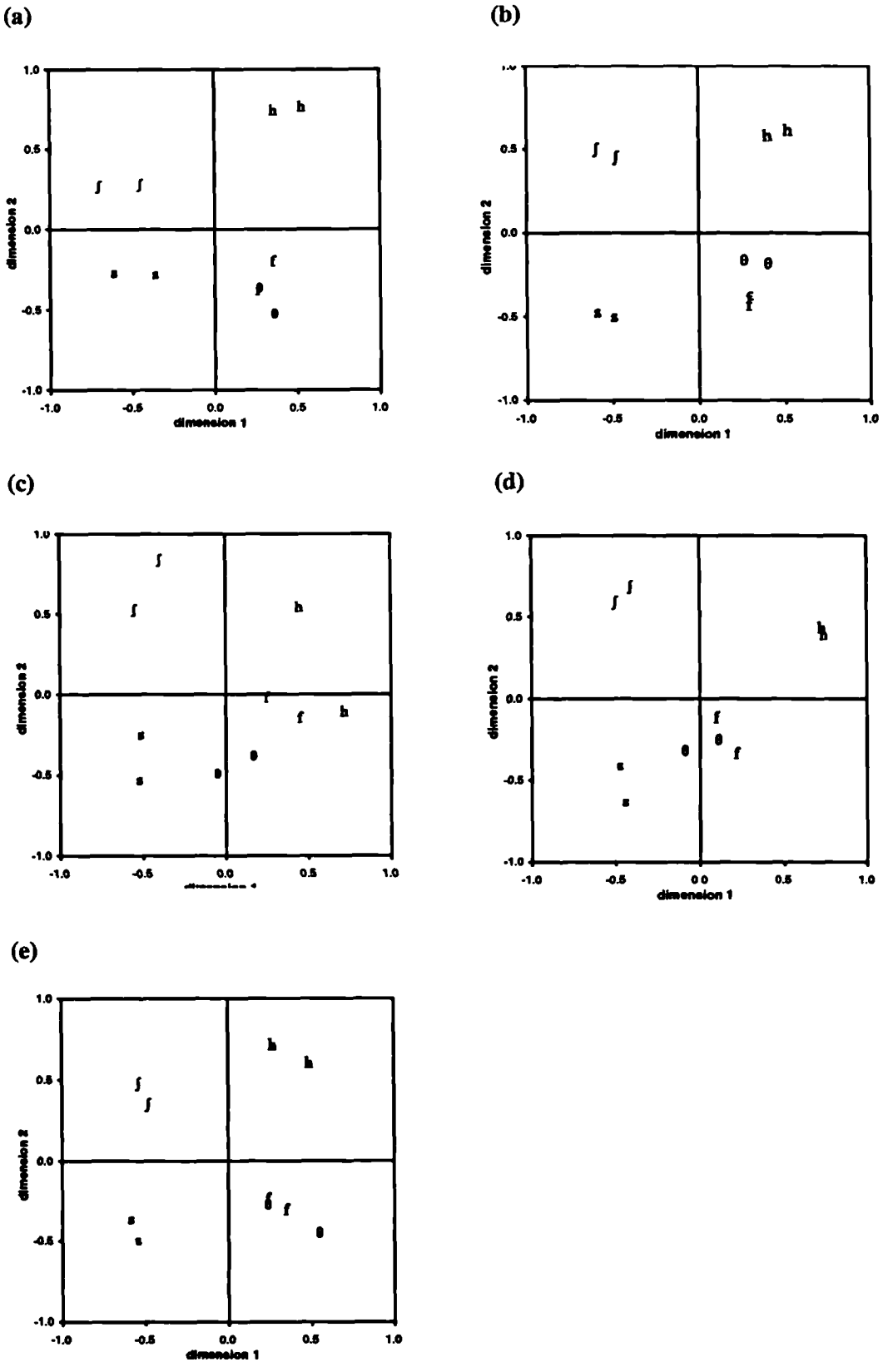



**Figure VI-4** Canonical correlation alignments between two productions for (a) speaker 1, (b)speaker 2, (c) speaker 3 (d) speaker 4 (e) speaker 5.

As was the case in the spectral variations, the spatial variations may also be due to RMS level differences. That is, the variations observed may be caused by differences in the overall spectral energy levels, rather than amounting to genuine differences between the spectral shapes of the fricatives.

To test this hypothesis, the overall RMS levels of each fricative were adjusted to a fixed level. The same procedures of auditory filter analysis and MDS were carried out. The mean variance accounted for in the 2-dimensional MDS solutions for each fricative set were 0.958 and 0.035 respectively. The coordinates of the MDS analyses were used for canonical correlation analyses. The correlation coefficients are shown in Table VI-4. The optimally matched coordinates of the two separate productions of each speaker are presented in Figures VI-5 (a) to VI-5 (b).

| Speakers | Dimensions | Correlation | Significance |
|---|---|---|---|
| 1 | 1 | 1.000 | 0.010 |
| | 2 | 0.903 | 0.111 |
| 2 | 1 | 0.999 | 0.013 |
| | 2 | 0.940 | 0.072 |
| 3 | 1 | 0.971 | 0.095 |
| | 2 | 0.954 | 0.057 |
| 4 | 1 | 0.994 | 0.026 |
| | 2 | 0.972 | 0.038 |
| 5 | 1 | 0.988 | 0.033 |
| | 2 | 0.982 | 0.026 |

**Table VI-4.** Canonical correlation coefficients between the two separate productions of the fricatives by the same speakers, after loudness normalisation.

The correlation coefficients are much higher for speakers 4 and 5, and especially, for speaker 3. The improvement in correlations is clearly visible in the spatial configurations in Figures VI-5 (c) to (e).

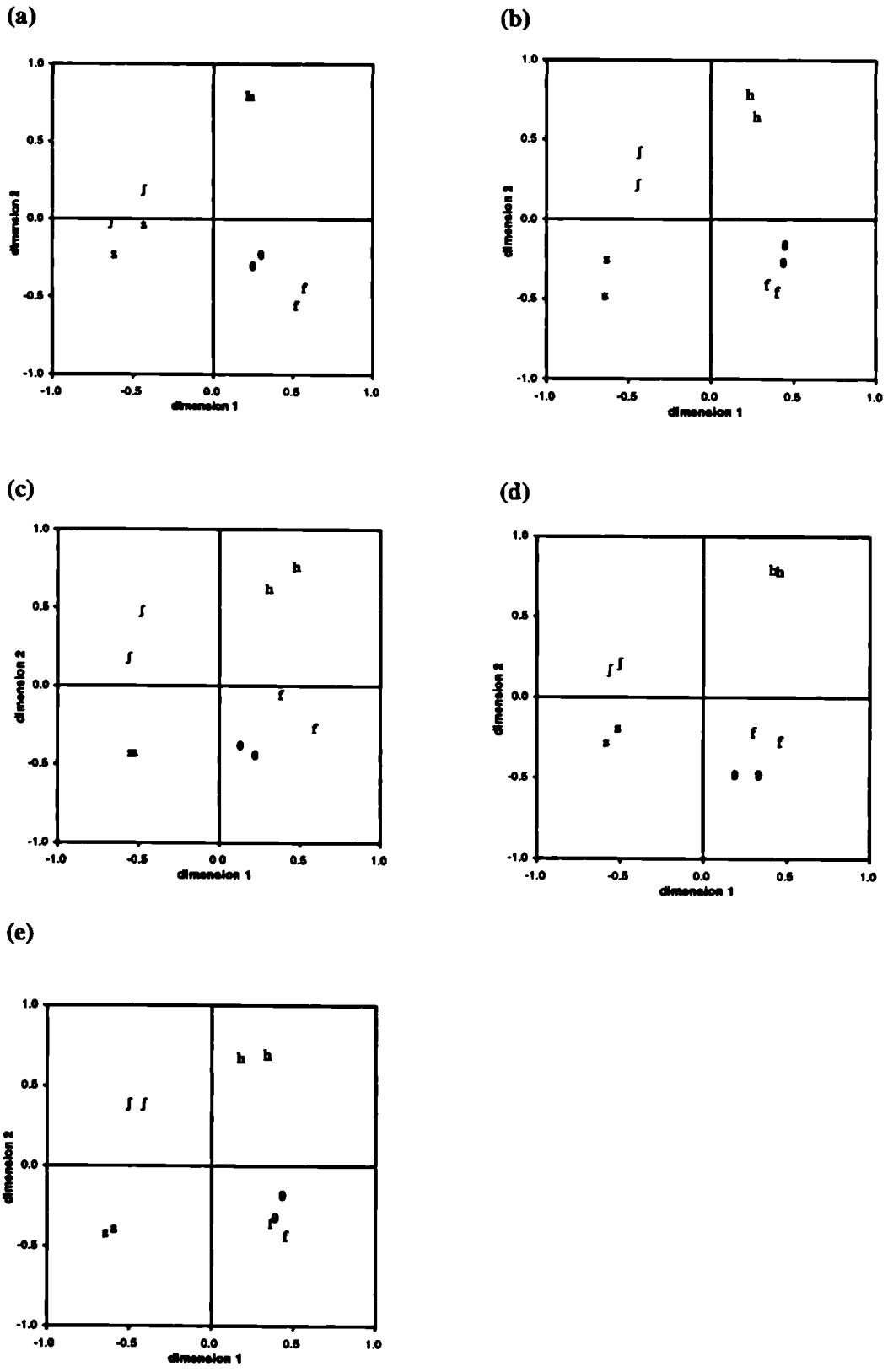



**Figure VI-5.** Canonical correlation scores between two productions for (a) speaker 1, (b) speaker 2, (c) speaker 3 (d) speaker 4 (e) speaker 5, after RMS level normalisation.

For speakers 1 and 2, the correlations have slightly decreased. However, by examining the spatial arrangements in Figures VI-5 (a) and (b), it can be shown that this is largely due to the shift in the relative positions of /s/ and /ʃ/; intensity normalisation has reduced the Euclidean distance between them. In fact, this same trend is observable for most of the configurations of these two fricatives.

Overall, the maps of two different productions (following normalisation) are much more stable, especially for speaker 3; we observe a considerable change in the distribution of the fricatives in Figure VI-5 (c), compared to the map in Figure VI-4 (c), before intensity normalisation.

### 4.2 Inter-speaker variations

In order to investigate inter-speaker variation, the auditory configurations of different speakers (in Figure VI-5) were plotted on the same axes. This is presented in Figure VI-6. This general auditory map of fricatives based on 10 productions shows that all the fricative regions are distinguishable from one another, and are roughly organised in terms of their 'place' and 'sibilance' properties (except that there are some overlapping areas for the fricatives /f/ and /θ/).

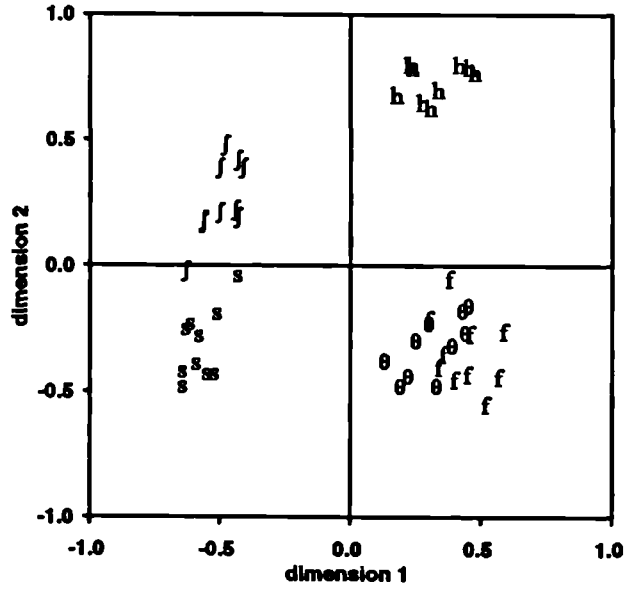



**Figure VI-6.** General auditory map of English fricatives based on 10 productions by 5 speakers.

This implies that inter-speaker variation is within the range of each fricative category in the auditory space, and thus, speaker-dependent variation is relatively small.

## 5 Auditory prototypes and acoustic correlates

Since speaker variation falls within an acceptable range, the results of different speakers can be pooled together to obtain an average auditory map. For this purpose, an average distance matrix was first calculated from each distance matrix of each fricative set. This average distance matrix was analysed by 2-way MDS. The resulting 2-dimensional solution is presented below.

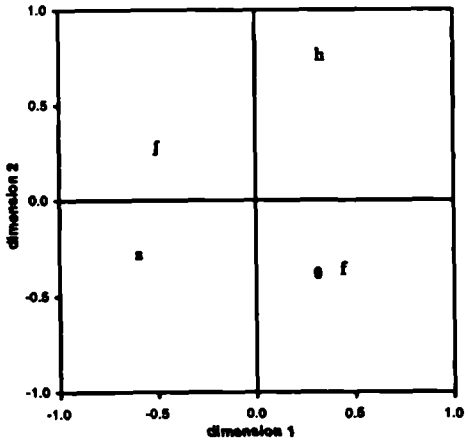



**Figure VI-7.** An average auditory map.

Each fricative point corresponds roughly to the centre of gravity of each fricative area in Figure VI-6, and thus, these points may be viewed as *auditory prototypes* (See §VII.3 for details). The place and sibilance properties are also maintained here.

At this point, it would be useful to make an informal comparison between the stimuli used in the perception tests and these prototypes. The most noticeable differences are that, in the Euclidean auditory space of the whole syllable stimuli (in Figure V-1 (a)), /f/ and /θ/ are much more distinct from each other, and the 'place' property was not clear. Thus, the stimuli in the perception tests may be regarded as acceptable variants of each fricative prototype, since they could be correctly identified, though this general auditory organisation of fricatives would not have been found from that particular set of tokens.

This fact must be taken into consideration in the overall interpretation of the relationship between perceptual and auditory spaces (see §VII.2).

Although the auditory dimensions were interpreted in terms of phonetic properties, the question of whether they may be related to any concrete physical properties of spectra is not investigated. The issue at this point is whether there are many acoustic parameters that correspond to each auditory dimension, or whether there may be a one-to-one correspondence between auditory and acoustic properties. In an attempt to relate the auditory dimensions to the acoustic properties of fricatives, the average spectrum for each fricative (in Figure VI-3) is placed on the auditory map (in Figure VI-6). This is shown in Figure VI-8.

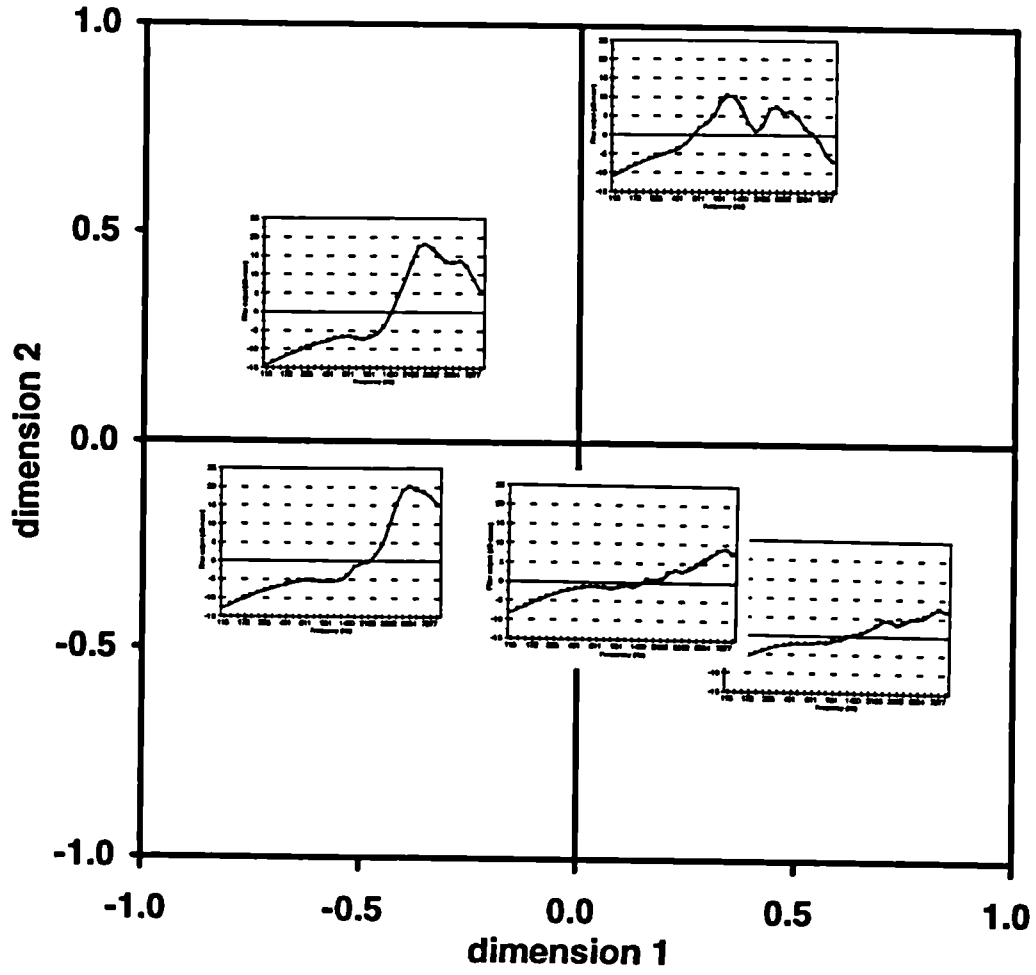



**Figure VI-8.** The average spectrum of each fricative is placed on the corresponding region of each fricative on the auditory axes.

The auditory dimension 1 may be related to 'peakiness' of spectra — the distance between the mean and main peak amplitudes, or centres of gravity of the spectra — while dimension 2 may be related to the frequencies of the peaks. The fact that each dimension in the general auditory map could also be related to a concrete physical property, as well as to phonetic and perceptual properties, may imply something quite fundamental about the speech processing mechanism, and an account of this result is provided in the next chapter.

## 6 Conclusion

The most remarkable result to emerge from the production tests is the general auditory map of fricatives in Figure VI-6; the map displays 'place' and 'sibilance' dimensions. Consequently, perceptual space — which has clear phonetic interpretations — must be well accounted for by the auditory map. This fact was not immediately apparent in the preceding chapter, since the stimuli involved in the perception tests were slightly different to the average auditory map of English fricatives.

The auditory dimensions were also closely related to peakiness and frequencies of peaks in the fricative spectra.

A fuller discussion on relationship between phonetic, perceptual, auditory, and acoustic properties is found in the next chapter (§VII.2).

Finally, the Euclidean metric — after RMS level adjustment and non-linear time alignments — is found to be quite effective in representing the auditory space of fricatives.

# *Chapter VII. Conclusion*

## 1 Introduction

The principal achievement of this research is that studies of spatial representations on vowels have been successfully replicated to a set of consonants, and furthermore, that the findings are congruent with those obtained in vowel studies. The results have shown a close relationship between perceptual and auditory spaces; and the phonetic and physical correlates of these spaces have also been identified. These results suggest a unified experimental paradigm in which the development of speech perception models may be investigated in parallel for both vowels and consonants in terms of spatial representations.

## 2 Main findings

### 2.1 General characteristics of spatial representations

The first main characteristic of the spatial representations discovered in the experiments was that two-dimensional solutions proved adequate in accounting for almost all the variation in the perceptual and auditory domains. This means that the parameters involved in perceptual similarity judgments and in spectral similarity of fricatives were reducible to two main components. Since there is a similar number of parameters involved in both domains, and also the correlation between perceptual and auditory spaces was high for fricatives and fricative-like sounds, this leads to the strong conjecture that the two domains may be closely related to each other. Therefore, the study of spatial representations enables us to identify key factors involved in each domain of the processing, and to demonstrate simple correlation across the different domains. This result stands in sharp contrast to the contemporary detailed cue studies (reviewed in §II.4) in which many different spectral characteristics seem to be intricately interwoven and often interact in specifying the perception of any one fricative category.

Secondly, the acoustic correlates of the auditory dimensions could be identified as *peakiness* (the difference between the mean and peak amplitudes) and frequency of main peak in the spectra (Figure VI-8). The spectra were obtained from the averages of

multiple speaker productions of fricatives, after critical bandpass filter bank analysis and non-linear intensity scaling. The amplitudes were normalised by subtracting the means. The issue of whether these spectral properties could vary according to vowel context was not investigated; however, peakiness and frequencies of peak are not expected to change. One obvious exception to this generalisation is /h/ on the peak frequency dimension, in that its peak locations are known to be sensitive to the quality of a following vowel. However, the highest peak frequency for /h/ corresponds to the presence of a following high vowel such as /i/ — and even in this case, the peak frequencies will be lower than those for /ʃ/ on dimension 2 in Figure VI-8. Thus, the acoustic interpretations for the auditory dimensions are likely to remain unaltered, although this remains to be confirmed in a later study (especially with regard to the effect of lip rounding).

The recurring properties in the perceptual and auditory spaces were the traditional phonetic characteristics of *place* and *sibilance* (Chapters IV and VI).

We have, therefore, made a clear illustration of how the perceptual and auditory domains of speech processing may be related to one another in terms of spatial representations; furthermore, these dimensions have clear phonetic interpretations, and are also related to concrete physical properties of fricative spectra. This is a novel finding, to the extent that spatial representations had never before been clearly established for consonants; as a consequence, the relationship between the spatial representations across the different domains had never been open to investigation. A limitation within the present study was that the spatial representations of phonetic (place and sibilance features) and acoustic domains were not formulated; they were merely referenced as possible correlates to the perceptual and auditory dimensions. A future spatial formulation of these domains may refine our understanding of each domain and the relationship across them.

### 2.2 Auditory modelling

Another important finding was that the auditory processing in fricative data was adequately modelled by the auditory transformations used in the vowel data (§V.2.2). A 1/3-octave bandpass filter bank analysis, coupled with non-linear intensity scaling, adequately modelled the essential peripheral perceptual processing. In order to account for the time-varying spectral properties in fricatives, a non-linear time alignment procedure

was employed. Auditory distances between fricatives were most accurately modelled by the Euclidean distance metric. A shortcoming of this metric is that it tends to be sensitive to the amplitude level adjustment (specifically, RMS values). A more detailed investigation into the effects of such adjustments in Euclidean distance modelling is required in the future. However, for the purpose of the present study, variations in amplitude differences could be effectively controlled by RMS level adjustments; the resulting auditory representations were then found to be congruent with spatial representations in other domains, as well as with the spectral properties of fricatives.

On the other hand, the Slope and N2D metrics, which were based on more selective models of spectral analysis, did not produce any reliable results (Chapter V), enhancing the credibility of the Euclidean metric. This result was attributed to specific materials used in designing these metrics in the previous studies (Klatt, 1982a; Assmann & Summerfield, 1989).

### 2.3 Variations in spatial relationship in different stimuli set

The relationships between phonetic, perceptual, and auditory spaces were expected to vary with the quality (degree of artificiality) of stimuli and listening modes (speech vs. nonspeech). The main perception tests were designed to investigate such potential variations. The stimuli consisted of a whole range of fricative sounds — from real fricative syllables to two-formant LPC synthesised fricatives — in order to allow the detection of any possible differences in the perceptual mechanism from real speech to evidently artificial tokens.

For the real fricatives, the perceptual dimensions were clearly related to the phonetic properties of fricatives. Excising the following transition and vowel sections had no discernible effect on the perceptual configurations. This implies that the perceptual judgments were based largely on the fricative part of the spectra. In addition to the lack of vowels and transitions in the stimuli of the cut-out set, the spectral shapes were also modified for duration adjustment (§V.2.1). Nevertheless, the changes were clearly within an acceptable perceptual *range* (§3) for each fricative category. A fuller understanding of such an acceptable perceptual range would require further research into the variations that may come about as a result of different surrounding contexts.

The phonetic interpretation of perceptual dimensions was also possible for the sets LPC22 and LPC10a for one subject group, but not for the other. This result was attributed to possible differences between the two listening groups in their modes of perception — that is, one group was listening in the speech mode of perception, and the other in nonspeech mode. However, the differences in perception between the two subject groups were very subtle, and any suggestion of a perceptual 'switch' from speech to nonspeech mode could only be confirmed in relation to the correlation between the perceptual and corresponding auditory maps. The initial hypothesis was that, if there were a perceptual 'switch' from the whole syllable set to the LPC4 set — which is indicated by the phonetic interpretability of the perceptual dimensions — then the stimuli with phonetically interpretable perceptual dimensions would be less well correlated to their corresponding auditory dimensions than other stimuli. As already alluded to, however, the correlations between the perceptual and auditory spaces were high for all the stimuli sets. This may be attributed to the fact that the general auditory space (Figure VI-6) was clearly related to phonetic properties; thus, the question of whether stimuli were perceived according to phonetic properties (in speech mode) or auditory properties may not indicate whether a particular set of stimuli was perceived in speech or nonspeech mode. It seems that the perception of both speech and nonspeech sounds is well accounted for by their auditory properties. The difference between speech and nonspeech perception processes lies in the fact that, for speech sounds, auditory organisation is clearly related to phonetic properties.

These observations have been made possible only by virtue of a new experimental design. The difference between speech and nonspeech perception lay beyond the scope of previous experimental designs, in which often specific aspects of spectra were manipulated to address a particular set of experimental questions (e.g. Sidwell & Summerfield, 1986; Krull, 1990).

## 2.4 Summary

In summary, all the main findings show that the results in the Pols *et al.* (1969) study are also confirmed in the case of fricative consonants. The perception of vowels could be mainly explained in terms of whole spectral auditory analyses. The main spatial dimensions

in the perceptual and auditory spaces were related to the traditional articulatory properties of vowels and to F1 and F2 values. Present research shows that these claims could be extended to encompass fricative consonants; the perceptual and auditory spaces have been found to correlate closely with corresponding phonetic and physical properties. The auditory modelling of fricatives was also found to be compatible with vowel sounds. The particular importance of such a finding lies in the fact that fricative sounds do not have obvious two-dimensionally measurable parameters, comparable to F1 and F2 in vowels. This implies that such an approach to investigations into speech perceptual processes need not be restricted to vowel sounds; indeed, they may be expanded to include a whole range of other speech sounds which have not been tested in this way before.

## 3 Implications for speech perception theory

In §I.1, we indicated that the central problem in speech perception is how an invariant percept (output) arises from a variable signal (input). Two major theories which have attempted to account for this are the motor theory and the theories of acoustic invariance (§I.2). In fact, most perceptual theories can be categorised as either a *strong articulatory theory* or a *strong auditory theory* (Nearey, 1991). Fowler's *direct realism* (1986) is another example of strong articulatory theory, while models of speech perception such as Klatt (1988) and Miller (1989) may be placed firmly in the strong auditory camp.

However, the results of the present experiment show that this kind of articulatory vs. auditory bifurcation is unnecessary. In fact, the *invariant properties* in a message are maintained throughout the various domains (§I.1) of speech processing. Although the medium of the message transfers changes from one domain to another, in each domain, essential parameters are kept in low dimensionality. In our fricative example, as well as in vowels, only two parameters are found to be essential even in the physical domain (spectral shapes), and clear correlates of these two parameters were found in the auditory, perceptual, and phonetic/articulatory domains.

Because of variability in the production and environmental conditions, we are compelled to accept a certain *range* of *variability* within any one category of sound in each domain. For example, in Figure VI-6, each fricative category occupies a distinct *region* rather than a single *point* on the auditory space. This acceptable range may be

defined through accumulated linguistic experience.

This line of argument is consistent with models of speech perception such as the *prototype theory* (Kuhl, 1995), according to which the correct identification of speech segments depends on the perceived distances between speech stimuli and a prototype in perceptual space. According to this theory, every human being is born with a universal perceptual space. Exposure to a particular language type determines the most representative instances of phonetic categories (prototypes), these functioning like *perceptual magnets* that attract similar sounds to the same category. These perceptual prototypes are also related to the formation of *articulatory targets* (Kuhl & Meltzoff, 1995). As with perception, infants are born with universal production space. This space becomes language-specific — by forming a tight clustering of each speech category (a kind of articulatory magnet) — as infants imitate sound patterns they hear. This result illustrates the close interaction between auditory and articulatory representations in early stages of development. However, most of the evidence has been drawn from infant perceptual and production spaces of vowel categories. The present results of fricative data in terms of spatial representations suggest an expansion of such a theory, to account for a wider range of speech data and spectral properties associated with them.

## 4 Future developments

Future development along these lines of research involves perceptual tests based on multiple speaker productions. This may allow for more accurate observations of the correlations between phonetic, perceptual and auditory dimensions. Also, the results obtained in this study need to be verified for voiced fricatives and fricatives in different vowel contexts. The vocalic excitation may have effect on the overall spectral shape of fricatives. The auditory transformation model applied here has the potential to be expanded to include the study of other consonant categories. If the simple relationship found in this study between each different space holds for all the types of speech sounds, we may be nearer to formulating a complete theory of speech perception, combining vowels and consonants.

# *References*

Assmann, P. F. & Summerfield, Q. (1989) Modelling the perception of concurrent vowels: Vowels with the same fundamental frequency. *Journal of the Acoustical Society of America* 85. 327-338.

Baker, R. & Rosen, S. (1993) Temporal information in consonant identification. *Speech, Hearing and Language; work in progress UCL* 7. 3-29.

Bakkum, M. J., Plomp, R. & Pols, L. C. W. (1993) Objective analysis versus subjective assessment of vowels pronounced by native, non-native, and deaf male speakers of Dutch. *Journal of the Acoustical Society of America* 94. 1989-2004.

Behrens, S. & Blumstein, S. E. (1988) On the role of the amplitude of the fricative noise in the perception of place of articulation in voiceless fricative consonants. *Journal of the Acoustical Society of America* 84. 861-867.

Bladon, R. A. & Lindblom, B. (1981) Modelling the judgment of vowel quality differences. *Journal of the Acoustical Society of America* 69. 1414-1422.

Blomberg, M., Carlson, R., Elenius, K. & Granstrom, B. (1986) Auditory models as front ends in speech-recognition systems. In Perkell, J. S. & Klatt, D. H. (eds). *Invariance and Variability in Speech Processes*. pp108-114. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Carlson, R. & Granstrom, B. (1979) Model prediction of vowel dissimilarity. *Speech Transmission Laboratory Quarterly Progress and Status Report* No. STL- QPSR 3-4 /1979, Royal Institute of Technology, Stockholm, Sweden. 84-104.

Carlson, R., Granstrom, B. & Klatt, D. (1979) Vowel perception: The relative perceptual salience of selected acoustic manipulations. *Speech Transmission Laboratory Quarterly Progress and Status Report* No. STL-QPSR 3- 4 /1979, Royal Institute of Technology, Stockholm, Sweden, 73-83.

Carroll, J. D. & Chang, J. J. (1970) Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart-Young' decomposition. *Psychometrika* 35. 283-319.

Clark, J. & Yallop, C. (1990) *An Introduction to Phonetics & Phonology.* Basil Blackwell, Cambridge, Massachusetts.

Delattre, P. C., Liberman, A. M. & Cooper, F. S. (1963) Formant transitions and loci as acoustic correlates of place of articulation in American fricatives. *Studia Linguistica* 16. 104-121.

Denes, P. B. & Pinson, E. N. (1993) *The Speech Chain: the Physics of biology of spoken language*, 2nd ed. W. H. Freeman and Company, New York.

Fowler, C. A. (1986) An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics* 14. 3-18.

Fox, R. A., Flege, J. E. & Munro, M. J. (1995) The perception of English and Spanish vowels by native English and Spanish listeners: A multidimensional scaling analysis. *Journal of the Acoustical Society of America* 97. 2540-2551.

Fox, R. A. (1983) Perceptual structure of monophthongs and diphthongs in English. *Language and Speech* 26. 21-60.

Gimson, A. C. (1989) *An introduction to the pronunciation of English*, 4th ed. Edward Arnold, London.

Glasberg, B. R. & Moore, B. C. J. (1990) Derivation of auditory filter shapes from notched-noise data. *Hearing Research* 47. 103-38.

Harman, H. H. (1967) *Modern Factor Analysis.* The University of Chicago Press, Chicago.

Harris, K. S. (1958) Cues for the discrimination of American English fricatives in spoken syllables. *Language and Speech* 1. 1-17.

Hedrick, M. S. & Ohde, R. N. (1993) Effect of relative amplitude of frication on perception of place of articulation. *Journal of the Acoustical Society of America* 94. 2005-2026.

Heinz, J. M. & Stevens K. N. (1961) On the Properties of Voiceless Fricative Consonants. *Journal of the Acoustical Society of America* 33. 589-596.

Hughes, G. W. & Halle, M. (1956) Spectral properties of fricative consonants. *Journal of the Acoustical Society of America* 28. 303-310.

Holmes, J. N. (1988) *Speech Synthesis and Recognition.* Van Nostrand Reinhold, UK.

Jones, D (1975) *An outline of English Phonetics*, 9th ed. Cambridge University Press, Cambridge.

Jongman, A. (1989) Duration of frication noise required for identification of English fricatives. *Journal of the Acoustical Society of America* 85. 1718-1725.

Joos, M. (1948) Acoustic Phonetics. *Language*, Monographs 23 (Suppl. 24).

Kewley-Port, D. & Atal, B. S. (1989) Perceptual differences between vowels located in a limited phonetic space. *Journal of the Acoustical Society of America* 85. 1726-1740.

Klatt, D. H. (1982a) Prediction of perceived phonetic distance from critical-band spectra : a first step. *Proc. ICASSP-82: IEEE International Conference on Acoustic, Speech and Signal Processing.* 1278-1281.

Klatt, D. H. (1982b) Speech processing strategies based on auditory models. In Carlson, R. & Granstrom, B. (eds). *The representation of Speech in the Peripheral Auditory System.* pp181-196. Elsevier Biomedical Press, Amsterdam.

Klatt, D. H. (1988) Review of selected models of speech perception. In Marslen-Wilson, W. (ed.) *Lexical representation and process.* pp169-226. MIT Press. Cambridge, Messachusetts.

Klein, W., Plomp, R. & Pols, L. C. W. (1970) Vowel spectra, vowel spaces and vowel identification. *Journal of the Acoustical Society of America* 48. 999-1009.

Krull, D. (1990) Relating acoustic properties to perceptual responses: A study of Swedish voiced stops. *Journal of the Acoustical Society of America* 88. 2557-2570.

Kruskal, J. B. (1964) Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 29. 1-27.

Kruskal, J. B. & Wish, M. (1978) *Multidimensional Scaling*, Sage University Paper series on Quantitative Applications in the Social Sciences, 07-011. Sage Publications, Beverly Hills and London.

Kuhl, P. K. (1995) Mechanisms of developmental change in speech and language. *Proc. ICPhS 95 Stockholm.* Vol. 2, 132-139.

Kuhl, P. K. & Meltzoff, A. N. (1995) Vocal learning in infants: Development of perceptual-motor links for speech. *Proc. ICPhS 95 Stockholm.* Vol 1, 146-149.

Ladefoged, P. & Maddieson, I. (1996) *The Sounds of the World's Languages.* Blackwell, Oxford.

Liberman, A. M. & Mattingly, I. G. (1985) The motor theory of speech perception

revised. *Cognition* 21. 1-36.

Lindblom, B. (1990) Explaining phonetic variation : A sketch of the H & H theory. In Hardcastle, W. J. & Marchal, A. (eds). *Speech Production and Speech Modelling.* pp 403-439. Dordrecht, Kluwer.

Lindblom, B. & Maddieson, I. (1988) Some effects of inventory size: Obstruents and sonorants. In Hyman, L. M. & Li, C. N. (eds). *Language, speech and mind studies in Honour of Victoria A. Fromkin.* pp62-78. Routledge, London.

Lindblom, B. (1978) Phonetic aspects of linguistic explanation. *Studia Linguistica* 31. 137-153

Mann, V. A. & Repp, B. H. (1980) Influence of vocalic context on perception of the [ʃ]-[s] distinction. *Perception & Psychophysics* 28 (3). 213-228.

Mann, V. & Soli, S. D. (1991) Perceptual order and the effect of vocalic context on fricative perception. *Perception & Psychophysics* 49 (5). 399-411.

Manrique, A. M. B. & Massone, M. I. (1981) Acoustic analysis and perception of Spanish fricative consonants. *Journal of the Acoustical Society of America* 69. 1145-1153.

Marslen-Wilson, W. D. & Tyler, L. K. (1980) The temporal structure of spoken language understanding. *Cognition* 8. 1-71.

McClelland, J. L. & Elman, J. L. (1986) Interactive processes in speech perception: The TRACE model. In Rumelhart, D. E. & McClelland, J. L. (eds) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* Vol 2, Chapter 15. MIT Press, Cambridge, Massachusetts.

Miller, J. D. (1989) Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America* 85. 2114-2134.

Miller, G. A. & Nicely, P. E. (1955) An Analysis of Perceptual Confusions among some English Consonants. *Journal of the Acoustical Society of America* 27. 338-352.

Nearey, T. M. (1991) Perception: automatic and cognitive processes. *Proc. of the Twelfth International Congress of Phonetic Sciences.* Vol. 1, pp40-49. Universite de Provence, Service des Publications, Aix-en-Provence.

Nocerino, N., Soong, F. K., Rabiner, L. R. & Klatt, D. H. (1985). Comparative Study of Several Distortion Measures for Speech Recognition. *IEEE Transitions. Acoustics, Speech, and Signal Processing.* 25-28

Ohala, J. J. (1979) The contribution of acoustic phonetics to phonology. In Lindblom, B. & Ohman, S. (eds). *Frontiers of speech communication research.* pp355-63. Academic Press, London.

Peterson, G. E. (1951) The phonetic value of vowels. *Language* 27. 541-53.

Pols, L. C. W., van der Kamp, L. J. Th. & Plomp, R. (1969). Perceptual and physical space of vowel sounds. *Journal of the Acoustical Society of America* 46. 456-467.

Rakerd, B. (1984) Vowels in consonantal context are perceived more linguistically than are isolated vowels: Evidence from an individual differences scaling study. *Perception & Psychophysics* 35. 123-136.

Rakerd, B. & Verbrugge, R. R. (1985) Linguistic and acoustic correlates of the perceptual structure found in an individual differences scaling study of vowels. *Journal of the Acoustical Society of America* 77. 296-301.

Repp, B. H. (1982) Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin* 92 (1). 81-110.

LaRiviere, C., Winitz, H. & Herriman, E. (1975) The distribution of perceptual cues in English prevocalic fricatives. *Journal of Speech and Hearing Research* 18. 613-622.

Rosen, S. & Fourcin, A. (1986) Frequency selectivity and the perception of speech. In Moore, B. C. J. (ed) *Frequency selectivity in Hearing.* Chapter 7. Academic Press, London

Rosen, S. & Howell, P. (1987) Auditory, articulatory, and learning explanations of categorical perception in speech. In Harnad, S. (ed) *Categorical perception.* pp113-160. Cambridge University Press, Cambridge.

Rosner, B. S. & Pickering, J. B. (1994) *Vowel perception and production.* Oxford Science Pub., Oxford.

Roucos, S & Wilgus, A. M. (1985) High quality time-scale modification for speech. *ICASSP-85: IEEE International Conference on Acoustic, Speech and Signal Processing* 2. 493-496

Sakoe, H. & Chiba, S. (1978) Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions. Acoustics, Speech, and Signal Processing.* 26. 43-49.

Schiffman, S. S., Reynolds, M. L. & Young, F. W. (1981) *Introduction to*

*Multidimensional Scaling; Theory, Methods, and Applications*. Academic Press, New York.

Shadle, C. H. (1990) Articulatory-acoustic relationships in fricative consonants. In Hardcastle, W. J. & Marchal, A. (eds). *Speech production and speech modelling.* pp189-209. Kluwer, Dordrecht, Netherlands.

Shepard, R. (1972) Psychological representation of speech sounds. In David, E. E. & Denes, P. B. (eds). *Human communication: a unified view.* pp67-113. McGraw-Hill, New York.

Sidwell, A. & Summerfield, Q. (1986) The auditory representation of symmetrical CVC syllables. *Speech Communication* 5. 283-297.

Simon, C. & Fourcin, A. (1978) Cross-language study of speech pattern learning. *Journal of the Acoustical Society of America.* 63. 925-935.

Singh, S. & Black, J. W. (1966) Study of Twenty-Six Intervocalic consonants as Spoken and recognized by Four Language Groups. *Journal of the Acoustical Society of America* 39. 372-387.

Singh, S. & Woods, G. (1970) Perceptual structure of 12 American English vowels. *Journal of the Acoustical Society of America* 49. 1861-1866 .

Singh, S., Woods, D. R. & Becker, G. M. (1972) Perceptual Structure of 22 Prevocalic English Consonants. *Journal of the Acoustical Society of America* 52. 1698-1713.

Soli, S. D., Arabie, P. & Carroll, J. D. (1986) Discrete representation of perceptual structure underlying consonant confusions. *Journal of the Acoustical Society of America* 79. 826-837.

Soli, S. D. & Arabie, P. (1979) Auditory versus phonetic accounts of observed confusions between consonant phonemes. *Journal of the Acoustical Society of America* 66. 46-59.

Stevens, K. N. (1985) Evidence for the role of acoustic boundaries in the perception of speech sounds. In Fromkin, V. A. (ed) *Phonetic Linguistics: Essays in Honour of Peter Ladefoged.* pp243-255. Academic Press, New York.

Stevens, K.N. & Blumstein, S. E. (1979) In variant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America* 64. 1358-68.

Stevens, S. S. & Volkmann, J. (1940) The relation of pitch to frequency: a revised scale. *American Journal of Psychology* 53, 329-53.

Sydral, A. K. & Gopal, H. S. (1986) A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America* 79. 1086-1100.

Sussman, H. M. (1991) The representation of stop consonants in three-dimensional acoustic space. *Phonetica* 48. 18-31.

Wilson, K. V. (1963) Multidimensional Analysis of confusions of English consonants. *American Journal of Psychology* 76. 89-95.

You, H-Y. (1979) An acoustical and perceptual study of English fricatives. *M.A. thesis. University of Edmonton.* Edmonton, Canada. (cited in Jongman, 1989).

Zwicker, E., Terhardt, E. & Paulus, E. (1979) Automatic speech recognition using psychoacoustic models. *Journal of the Acoustical Society of America* 65. 487-498.

# *Appendix. Raw similarity matrices for each listener*

**Whole syllable set**

| **Subject 1** | **Subject 6** | **Subject 11** | **Subject 16** |
|---|---|---|---|
| 034201 | 032221 | 031123 | 041122 |
| 402031 | 401212 | 301024 | 402211 |
| 230311 | 200314 | 210421 | 120430 |
| 113041 | 123022 | 023032 | 114031 |
| 221401 | 111304 | 232003 | 120304 |
| 142120 | 031330 | 231040 | 202240 |
| **Subject 2** | **Subject 7** | **Subject 12** | **Subject 17** |
| 042013 | 042130 | 043102 | 041023 |
| 403021 | 401221 | 203104 | 301132 |
| 110341 | 110431 | 240301 | 220402 |
| 113041 | 023041 | 123031 | 102043 |
| 302302 | 013303 | 131302 | 210304 |
| 231040 | 111340 | 123400 | 210340 |
| **Subject 3** | **Subject 8** | **Subject 13** | **Subject 18** |
| 033112 | 042211 | 041122 | 040123 |
| 403201 | 303310 | 401203 | 401212 |
| 240112 | 220411 | 020341 | 020332 |
| 432001 | 232021 | 022042 | 321022 |
| 214201 | 032104 | 120304 | 221104 |
| 413200 | 013240 | 120340 | 311140 |
| **Subject 4** | **Subject 9** | **Subject 14** | **Subject 19** |
| 024220 | 043111 | 022033 | 041032 |
| 204130 | 401212 | 201034 | 400132 |
| 330310 | 220213 | 310420 | 130231 |
| 132031 | 203023 | 013033 | 312040 |
| 221203 | 210403 | 131104 | 311302 |
| 111340 | 131230 | 231130 | 230140 |
| **Subject 5** | **Subject 10** | **Subject 15** | **Subject 20** |
| 041122 | 043201 | 041023 | 041023 |
| 403111 | 401203 | 401023 | 402022 |
| 130321 | 230401 | 110422 | 210142 |
| 023041 | 143011 | 131032 | 112042 |
| 311104 | 013204 | 131104 | 021304 |
| 320140 | 012340 | 320140 | 230140 |

**No-transition set**

| **Subject 1** | **Subject 6** | **Subject 11** | **Subject 16** |
|---|---|---|---|
| 043120 | 030322 | 041122 | 040132 |
| 403021 | 302014 | 402121 | 401122 |
| 400123 | 220222 | 100342 | 110431 |
| 112042 | 223021 | 013042 | 024022 |
| 013303 | 233200 | 202303 | 124102 |
| 023320 | 102430 | 321040 | 122140 |
| **Subject 2** | **Subject 7** | **Subject 12** | **Subject 17** |
| 032122 | 041230 | 042013 | 030322 |
| 402013 | 402031 | 402103 | 401221 |
| 110242 | 200431 | 210313 | 300322 |
| 012043 | 024031 | 303031 | 111043 |
| 113203 | 133201 | 201403 | 230203 |
| 410230 | 112420 | 230230 | 212230 |
| **Subject 3** | **Subject 8** | **Subject 13** | **Subject 18** |
| 024013 | 034111 | 040132 | 040231 |
| 204013 | 401302 | 400132 | 401221 |
| 340111 | 310222 | 110341 | 220420 |
| 312022 | 232021 | 201043 | 113032 |
| 241201 | 122302 | 320104 | 120304 |
| 223210 | 220330 | 310240 | 211240 |
| **Subject 4** | **Subject 9** | **Subject 14** | **Subject 19** |
| 041122 | 042013 | 031123 | 042031 |
| 402031 | 403201 | 402022 | 401113 |
| 320032 | 310321 | 320311 | 400222 |
| 221041 | 024031 | 212023 | 322030 |
| 212203 | 130204 | 320104 | 311203 |
| 222130 | 131230 | 423010 | 220240 |
| **Subject 5** | **Subject 10** | **Subject 15** | **Subject 20** |
| 041023 | 043201 | 041131 | 040132 |
| 401023 | 402211 | 400123 | 401023 |
| 030232 | 420301 | 230113 | 100432 |
| 212032 | 323011 | 222031 | 103042 |
| 113104 | 113203 | 320104 | 302302 |
| 331030 | 212140 | 420130 | 121240 |

**Cut-out set**

| **Subject 1** | **Subject 6** | **Subject 11** | **Subject 16** |
|---|---|---|---|
| 031231 | 043210 | 040123 | 041023 |
| 403021 | 304102 | 401032 | 401113 |
| 130330 | 230320 | 230311 | 330220 |
| 124030 | 411031 | 213031 | 102043 |
| 042301 | 422002 | 131203 | 201403 |
| 023140 | 220240 | 311230 | 321040 |
| **Subject 2** | **Subject 7** | **Subject 12** | **Subject 17** |
| 032230 | 042301 | 022231 | 041122 |
| 403012 | 403021 | 403111 | 401221 |
| 330130 | 230401 | 340201 | 420202 |
| 212023 | 224011 | 224011 | 203023 |
| 231004 | 330103 | 320104 | 221104 |
| 231130 | 231040 | 320140 | 330130 |
| **Subject 3** | **Subject 8** | **Subject 13** | **Subject 18** |
| 042310 | 034120 | 040321 | 040132 |
| 303130 | 103420 | 402130 | 403012 |
| 130312 | 310231 | 220420 | 200332 |
| 232030 | 223012 | 212023 | 112033 |
| 212203 | 112204 | 112204 | 320203 |
| 211330 | 113320 | 211240 | 420130 |
| **Subject 4** | **Subject 9** | **Subject 14** | **Subject 19** |
| 023122 | 043120 | 032113 | 040321 |
| 303310 | 402121 | 302311 | 402310 |
| 420121 | 340120 | 210430 | 310420 |
| 413020 | 214030 | 223030 | 314020 |
| 142102 | 320302 | 033103 | 230104 |
| 413200 | 230320 | 231130 | 210340 |
| **Subject 5** | **Subject 10** | **Subject 15** | **Subject 20** |
| 042130 | 031411 | 040132 | 041023 |
| 403021 | 401113 | 402031 | 402112 |
| 230410 | 120232 | 340120 | 430201 |
| 124030 | 023023 | 113032 | 233011 |
| 320104 | 211204 | 231004 | 211204 |
| 320230 | 220330 | 230140 | 012340 |

**LPC22 set**

| **Subject 1** | **Subject 6** | **Subject 11** | **Subject 16** |
|---|---|---|---|
| 033121 | 041032 | 040222 | 030214 |
| 402121 | 304021 | 403111 | 303031 |
| 210340 | 220231 | 210322 | 340120 |
| 034021 | 323011 | 302041 | 301042 |
| 313201 | 223201 | 231103 | 201304 |
| 211240 | 320230 | 320230 | 310240 |
| **Subject 2** | **Subject 7** | **Subject 12** | **Subject 17** |
| 041032 | 030232 | 031132 | 031141 |
| 402031 | 401212 | 302122 | 402022 |
| 320122 | 220402 | 240310 | 330121 |
| 012034 | 323020 | 202033 | 222022 |
| 132004 | 421003 | 220204 | 102304 |
| 321040 | 111340 | 320140 | 221140 |
| **Subject 3** | **Subject 8** | **Subject 13** | **Subject 18** |
| 033301 | 031114 | 032032 | 031123 |
| 402211 | 103411 | 401023 | 402022 |
| 400321 | 220330 | 220240 | 230410 |
| 124021 | 223030 | 012043 | 312022 |
| 121204 | 322003 | 103204 | 420103 |
| 103330 | 213310 | 241030 | 320140 |
| **Subject 4** | **Subject 9** | **Subject 14** | **Subject 19** |
| 042121 | 041032 | 021232 | 031132 |
| 102232 | 403210 | 301222 | 402031 |
| 240121 | 420310 | 030322 | 230401 |
| 324001 | 214030 | 121042 | 422020 |
| 323200 | 232102 | 201304 | 211204 |
| 412120 | 320140 | 410230 | 420130 |
| **Subject 5** | **Subject 10** | **Subject 15** | **Subject 20** |
| 031222 | 032122 | 042013 | 041122 |
| 400123 | 203221 | 303022 | 402022 |
| 230311 | 230221 | 230302 | 120412 |
| 323020 | 113023 | 213031 | 324001 |
| 413101 | 103204 | 231004 | 221203 |
| 420130 | 111340 | 321130 | 221140 |

**LPC10a set**

| **Subject 1** | **Subject 6** | **Subject 11** | **Subject 16** |
|---|---|---|---|
| 041023 | 042130 | 040123 | 042103 |
| 301321 | 401230 | 400132 | 401023 |
| 210340 | 320230 | 110233 | 220114 |
| 023041 | 132022 | 221041 | 112042 |
| 012304 | 201304 | 331003 | 311401 |
| 123220 | 211240 | 241030 | 321130 |
| **Subject 2** | **Subject 7** | **Subject 12** | **Subject 17** |
| 031033 | 042211 | 012313 | 040123 |
| 401023 | 203131 | 203311 | 401023 |
| 130132 | 420220 | 030322 | 120421 |
| 311032 | 143020 | 202033 | 242020 |
| 231103 | 111403 | 021304 | 310204 |
| 221140 | 213040 | 310240 | 411130 |
| **Subject 3** | **Subject 8** | **Subject 13** | **Subject 18** |
| 031033 | 041113 | 040123 | 041221 |
| 401023 | 401113 | 400123 | 401023 |
| 130132 | 320401 | 010243 | 310312 |
| 311032 | 213013 | 012043 | 212032 |
| 231103 | 132004 | 211204 | 331003 |
| 221140 | 221230 | 320140 | 421120 |
| **Subject 4** | **Subject 9** | **Subject 14** | **Subject 19** |
| 022222 | 011314 | 040231 | 040132 |
| 204121 | 102223 | 400231 | 400123 |
| 430201 | 330103 | 330022 | 300232 |
| 332020 | 302023 | 103033 | 022033 |
| 333001 | 211204 | 331003 | 420103 |
| 224110 | 312310 | 231040 | 320140 |
| **Subject 5** | **Subject 10** | **Subject 15** | **Subject 20** |
| 040132 | 042220 | 032014 | 040123 |
| 400123 | 204112 | 402112 | 400132 |
| 130330 | 030313 | 340201 | 330211 |
| 121042 | 213022 | 220024 | 312022 |
| 321103 | 213103 | 210403 | 430102 |
| 420130 | 113230 | 323110 | 420130 |

**LPC10 set**

| **Subject 1** | **Subject 6** | **Subject 11** | **Subject 16** |
|---|---|---|---|
| 041122 | 030232 | 031033 | 030124 |
| 402121 | 402121 | 303022 | 402013 |
| 120430 | 110431 | 240022 | 120421 |
| 102034 | 224011 | 302023 | 302023 |
| 332002 | 312202 | 130204 | 210403 |
| 130240 | 221140 | 240220 | 410230 |
| **Subject 2** | **Subject 7** | **Subject 12** | **Subject 17** |
| 031132 | 041212 | 042022 | 041113 |
| 401023 | 401122 | 401212 | 401122 |
| 230230 | 130303 | 130231 | 220312 |
| 121033 | 233002 | 122023 | 132022 |
| 411202 | 421003 | 201304 | 231103 |
| 221140 | 430210 | 311230 | 420130 |
| **Subject 3** | **Subject 8** | **Subject 13** | **Subject 18** |
| 041212 | 032104 | 040132 | 030412 |
| 403012 | 203140 | 401023 | 401122 |
| 230320 | 030421 | 020224 | 120313 |
| 034021 | 233011 | 012034 | 134020 |
| 322201 | 430102 | 320203 | 111304 |
| 131230 | 321220 | 420130 | 420310 |
| **Subject 4** | **Subject 9** | **Subject 14** | **Subject 19** |
| 022411 | 042112 | 030214 | 021124 |
| 403210 | 403021 | 302032 | 401122 |
| 420310 | 420220 | 230212 | 240301 |
| 114031 | 423010 | 223012 | 220033 |
| 132301 | 420301 | 340102 | 310204 |
| 413200 | 320320 | 340120 | 301240 |
| **Subject 5** | **Subject 10** | **Subject 15** | **Subject 20** |
| 030133 | 023203 | 032113 | 040132 |
| 303121 | 202222 | 304201 | 302041 |
| 230230 | 130420 | 340210 | 120322 |
| 122032 | 233020 | 410023 | 113041 |
| 421102 | 221203 | 210403 | 022204 |
| 430120 | 302320 | 320320 | 201430 |

**LPC4 set**

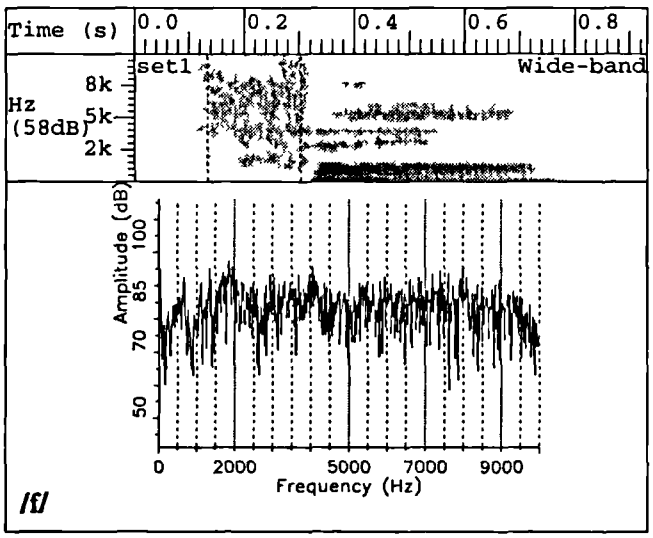
| **Subject 1** | **Subject 6** | **Subject 11** | **Subject 16** |
|---|---|---|---|
| 022213 | 033121 | 041221 | 010234 |
| 302221 | 303112 | 302032 | 303022 |
| 110422 | 220321 | 200332 | 240220 |
| 112033 | 411031 | 222031 | 202042 |
| 320302 | 331102 | 231004 | 210304 |
| 211330 | 412030 | 341020 | 401230 |
| **Subject 2** | **Subject 7** | **Subject 12** | **Subject 17** |
| 040132 | 030142 | 010324 | 030133 |
| 401032 | 300232 | 203311 | 300124 |
| 200323 | 230230 | 110341 | 210331 |
| 211033 | 202033 | 122032 | 312031 |
| 341101 | 240103 | 301204 | 410203 |
| 330130 | 430210 | 321130 | 331120 |
| **Subject 3** | **Subject 8** | **Subject 13** | **Subject 18** |
| 043102 | 022222 | 011134 | 021232 |
| 202141 | 303031 | 202024 | 401122 |
| 100423 | 020431 | 230311 | 210331 |
| 114022 | 214030 | 012034 | 203041 |
| 321202 | 123103 | 320203 | 103204 |
| 233020 | 201340 | 310330 | 310330 |
| **Subject 4** | **Subject 9** | **Subject 14** | **Subject 19** |
| 020422 | 031024 | 030133 | 020224 |
| 104203 | 402022 | 104032 | 301312 |
| 130231 | 230320 | 320311 | 030421 |
| 132013 | 432010 | 301042 | 202042 |
| 123202 | 330103 | 410203 | 310204 |
| 221320 | 430120 | 320140 | 410230 |
| **Subject 5** | **Subject 10** | **Subject 15** | **Subject 20** |
| 030214 | 022033 | 020224 | 020143 |
| 200233 | 202222 | 204013 | 400132 |
| 320410 | 200422 | 300421 | 220411 |
| 213040 | 204022 | 213031 | 013042 |
| 320203 | 313201 | 220204 | 222103 |
| 321040 | 312040 | 230140 | 420130 |

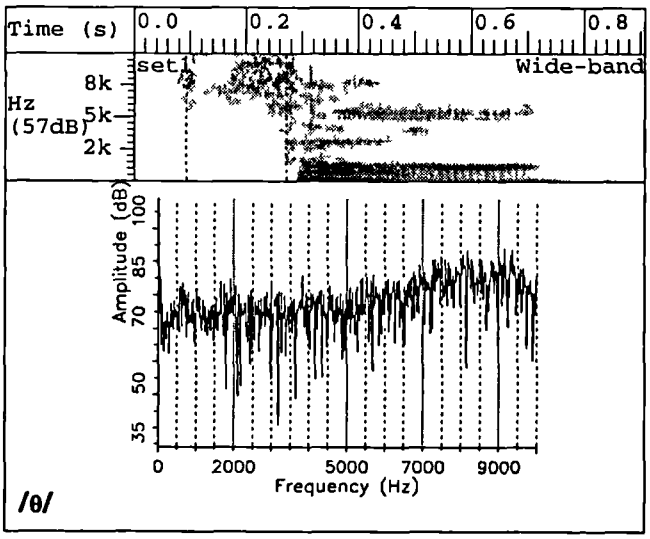
# *Addendum. Spectrographic and spectral analyses of the stimuli in Chapter IV.*

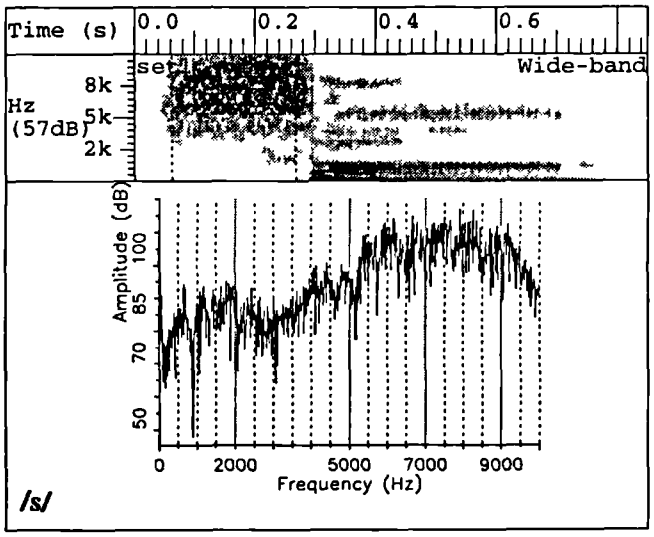
---

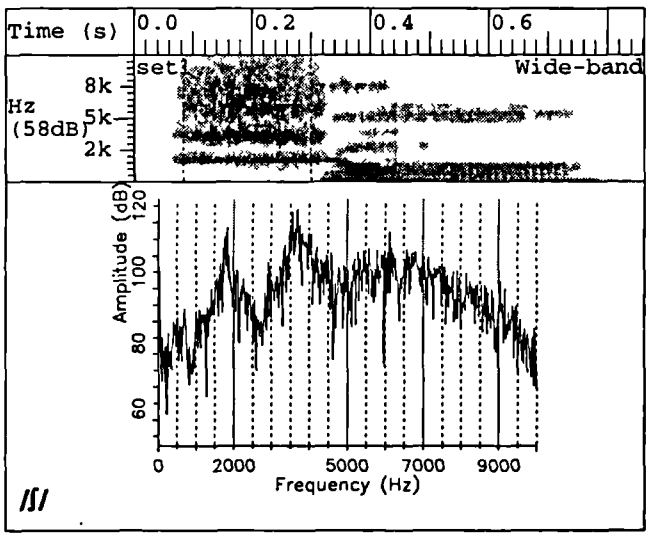
In the following pages, the spectrograms and average cross-sectional spectra of fricatives used in Chapter IV are presented. The spectra were obtained by FFT analyses of the marked regions on the respective spectrograms.

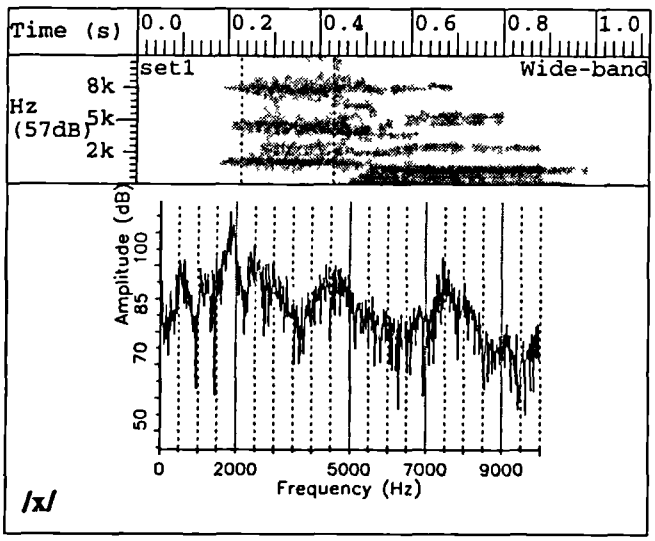
**Whole syllable set**

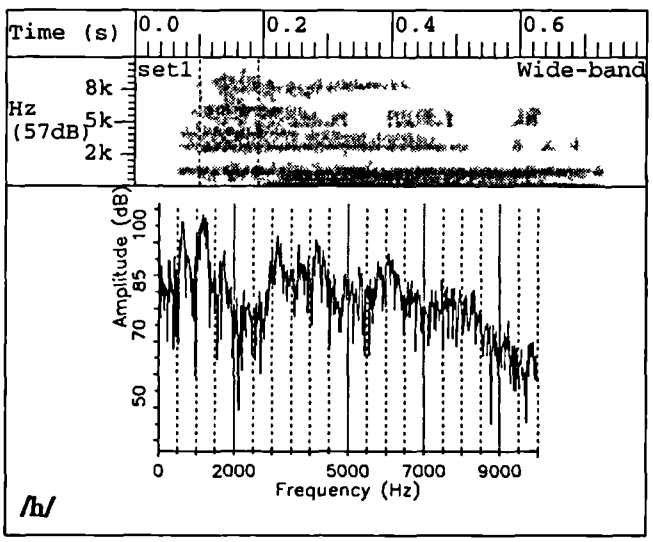


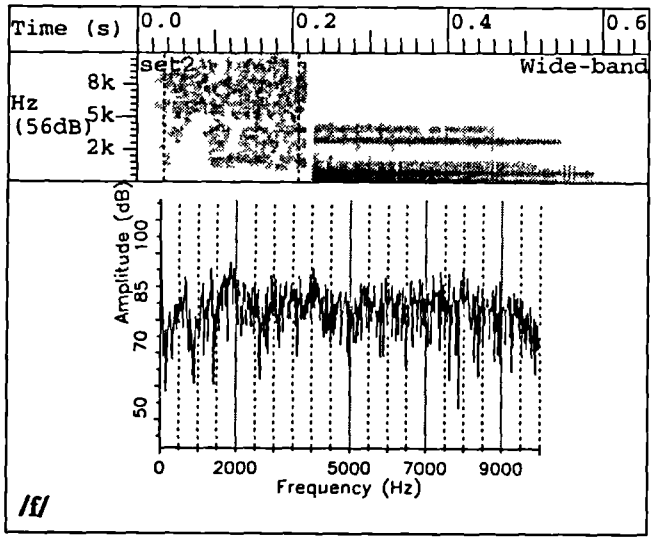



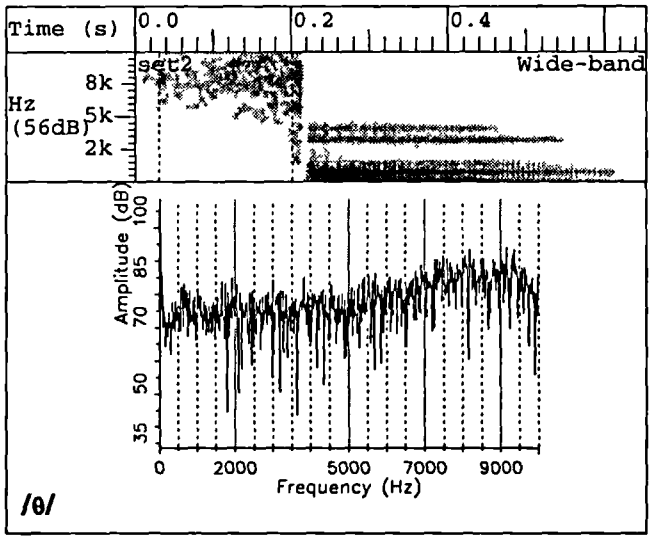



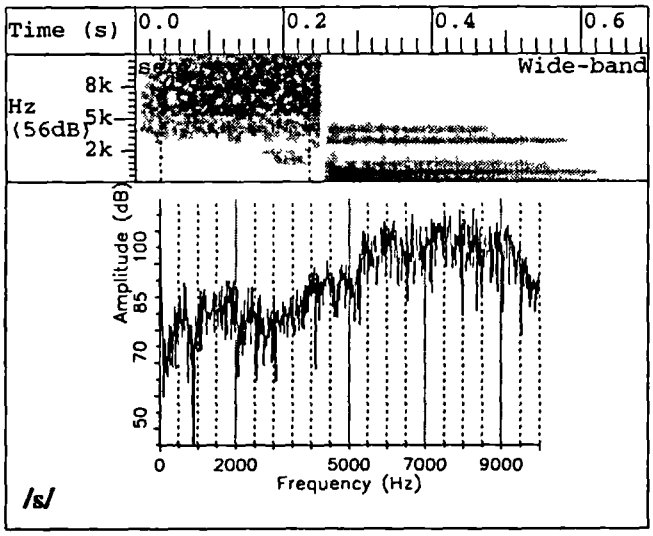



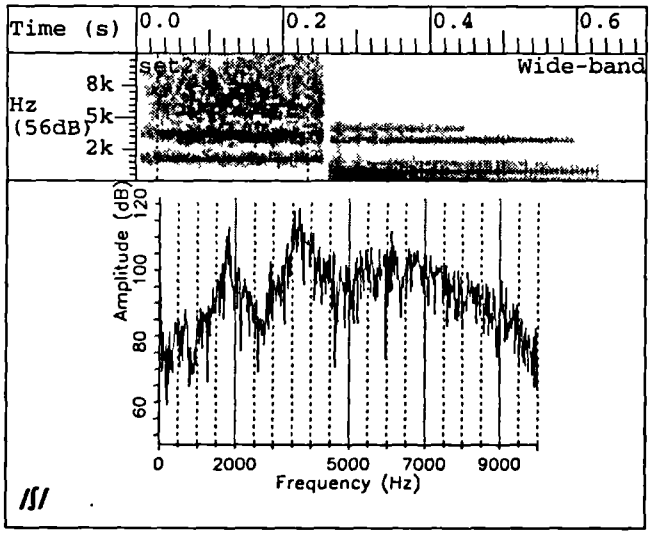



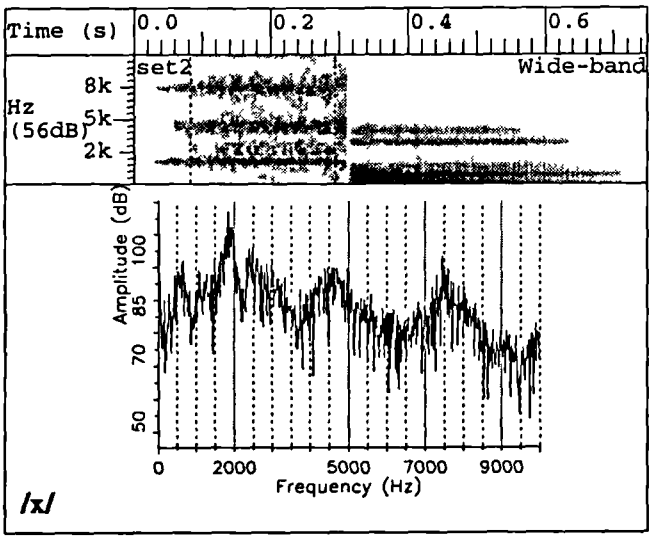
## No-transition set

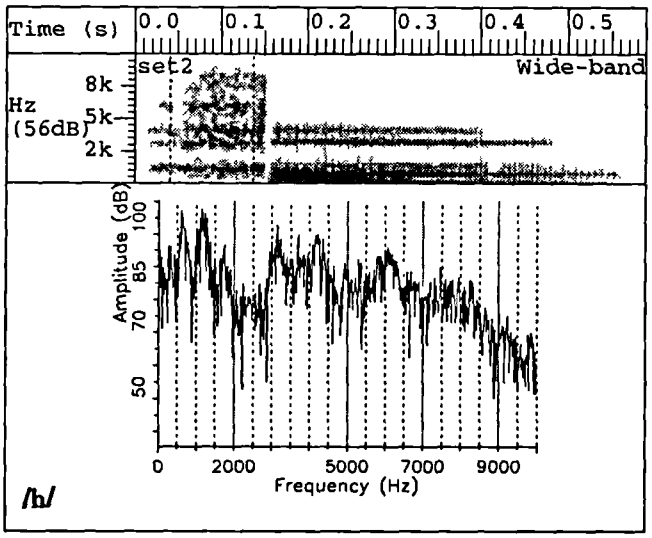


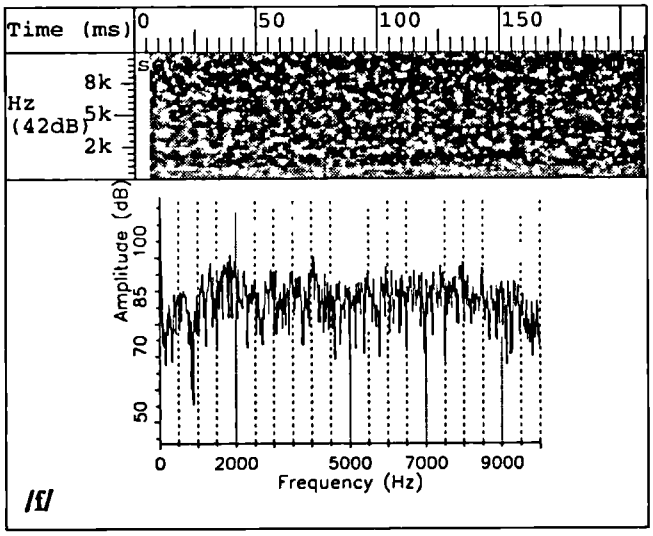



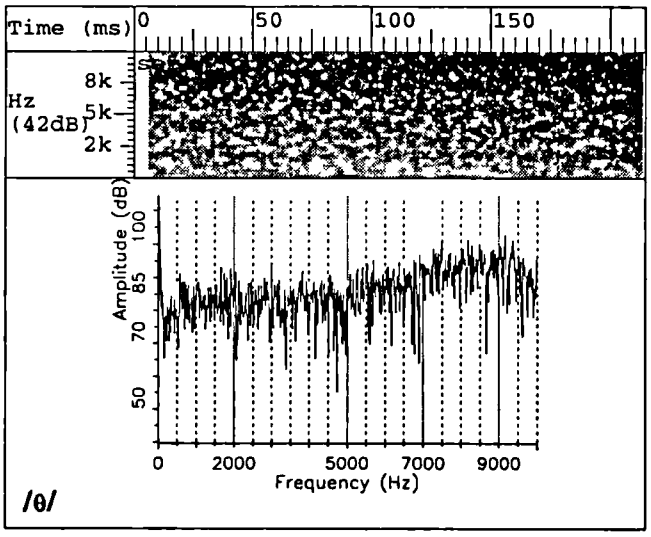



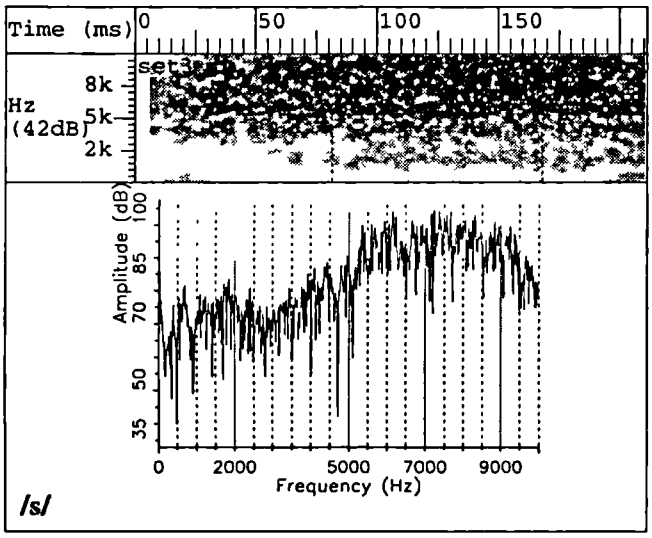



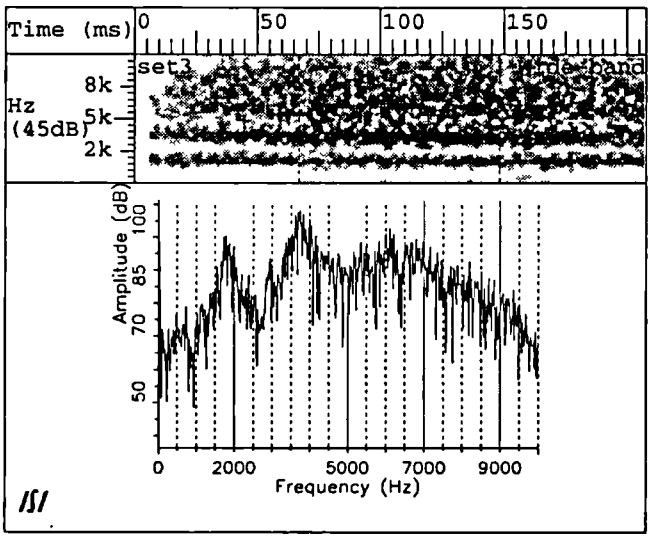



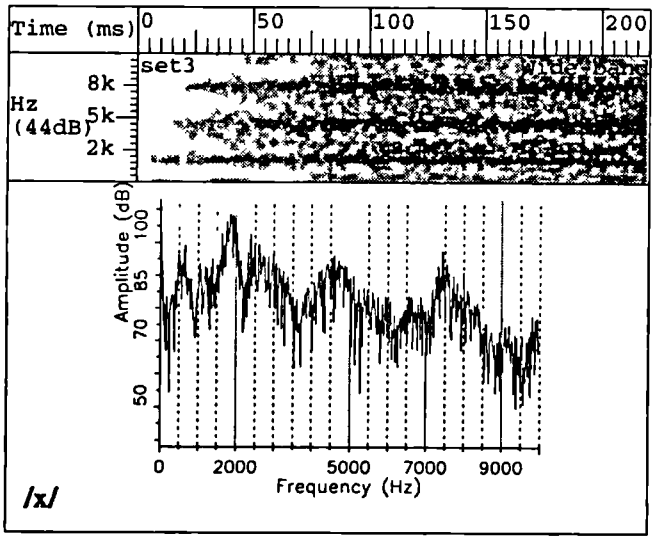
## Cut-out set

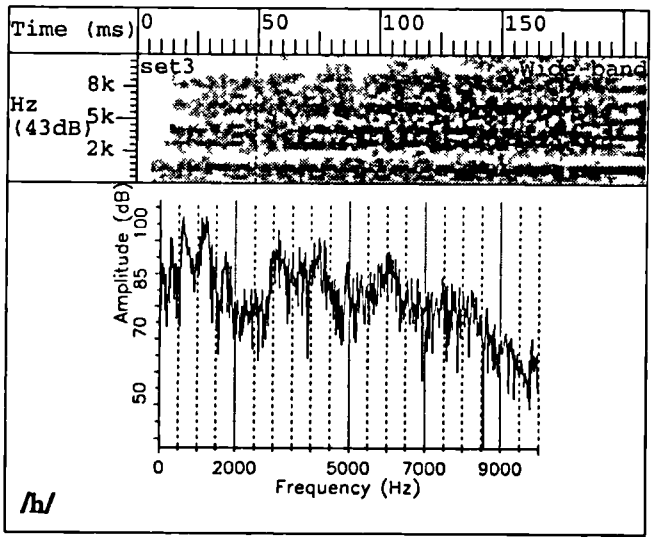

Time (ms)
0
50
100
150
Hz
(42dB)
8k
5k
2k
Amplitude (dB)
100
85
70
50
0
2000
5000
7000
9000
Frequency (Hz)
/ʃ/


Time (ms)
0
50
100
150
Hz
(42dB)
8k
5k
2k
Amplitude (dB)
100
85
70
50
0
2000
5000
7000
9000
Frequency (Hz)
/θ/

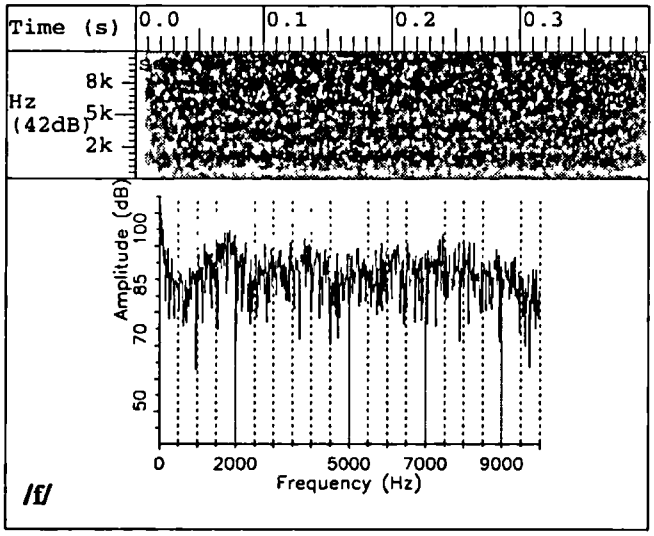

Time (ms)
0
50
100
150
set3
Hz
(42dB)
8k
5k
2k
Amplitude (dB)
100
85
70
50
35
0
2000
5000
7000
9000
Frequency (Hz)
/s/

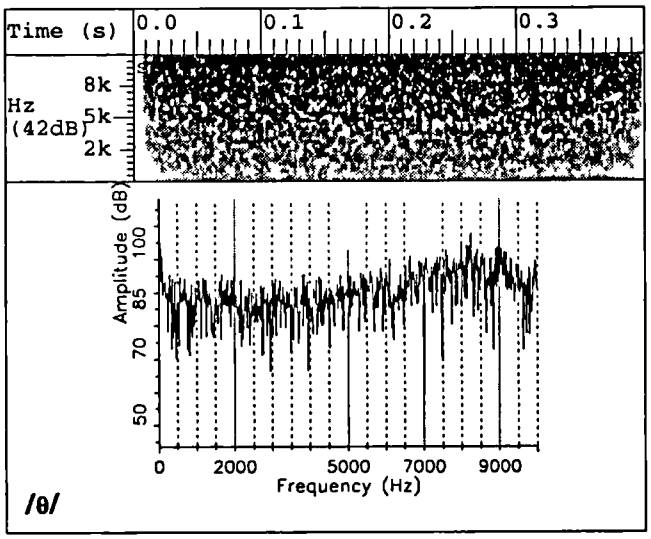

Time (ms)
0
50
100
150
set3
Wide band
Hz
(45dB)
8k
5k
2k
Amplitude (dB)
100
85
70
50
0
2000
5000
7000
9000
Frequency (Hz)
/ʃ/

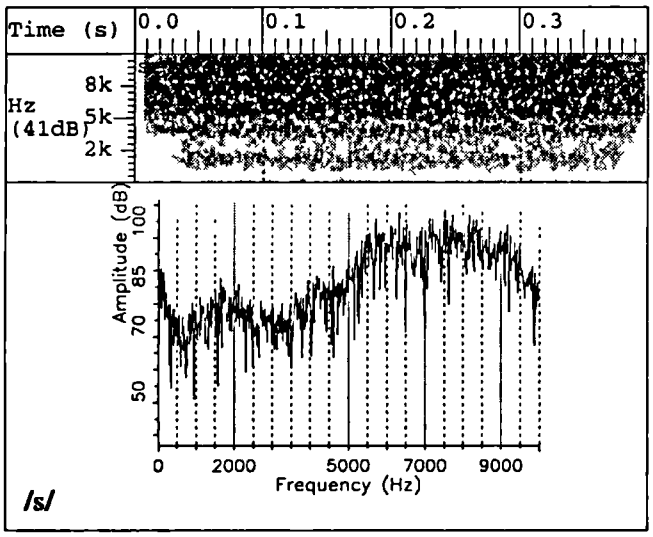

Time (ms)
0
50
100
150
200
set3
Wide band
Hz
(44dB)
8k
5k
2k
Amplitude (dB)
100
85
70
50
0
2000
5000
7000
9000
Frequency (Hz)
/x/

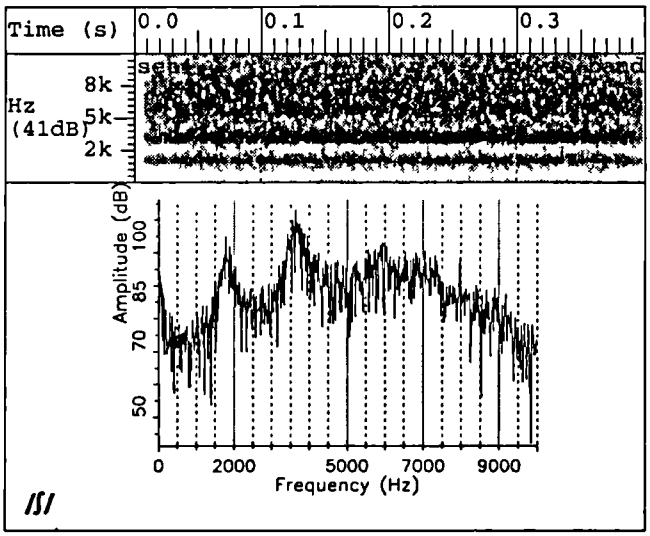

Time (ms)
0
50
100
150
set3
Wide band
Hz
(43dB)
8k
5k
2k
Amplitude (dB)
100
85
70
50
0
2000
5000
7000
9000
Frequency (Hz)
/h/

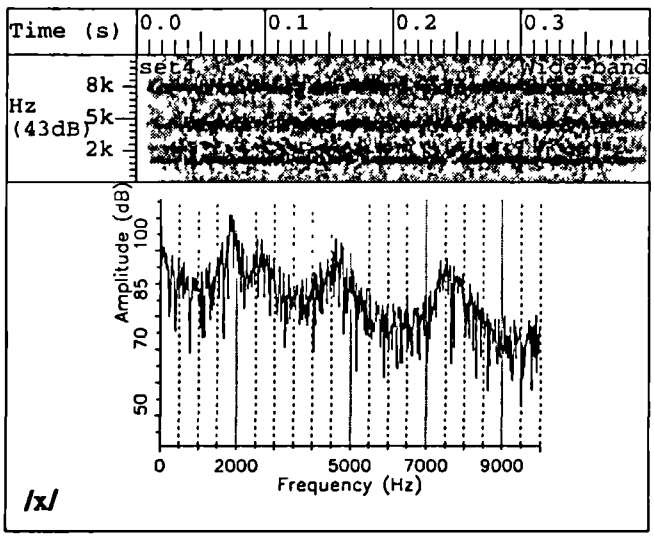
**LPC22 set**

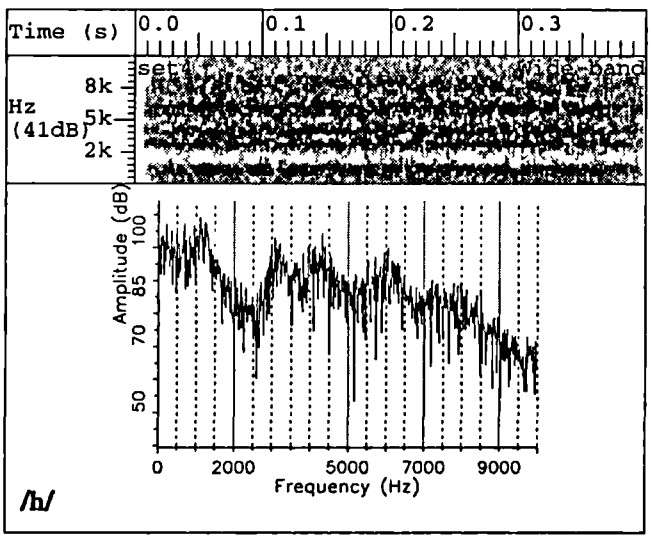

Time (s)
0.0
0.1
0.2
0.3
Hz
(42dB)
8k
5k
2k
Amplitude (dB)
100
85
70
50
0
2000
5000
7000
9000
Frequency (Hz)
/f/


Time (s)
0.0
0.1
0.2
0.3
Hz
(42dB)
8k
5k
2k
Amplitude (dB)
100
85
70
50
0
2000
5000
7000
9000
Frequency (Hz)
/θ/

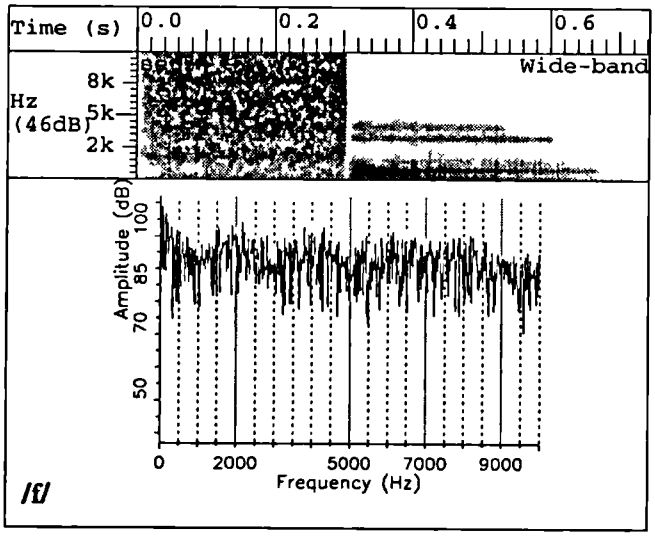

Time (s)
0.0
0.1
0.2
0.3
Hz
(41dB)
8k
5k
2k
Amplitude (dB)
100
85
70
50
0
2000
5000
7000
9000
Frequency (Hz)
/s/

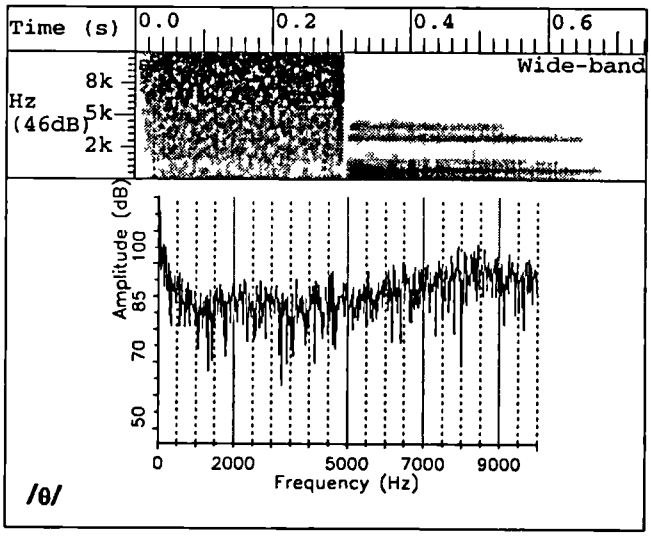

Time (s)
0.0
0.1
0.2
0.3
Hz
(41dB)
8k
5k
2k
Amplitude (dB)
100
85
70
50
0
2000
5000
7000
9000
Frequency (Hz)
/ʃ/

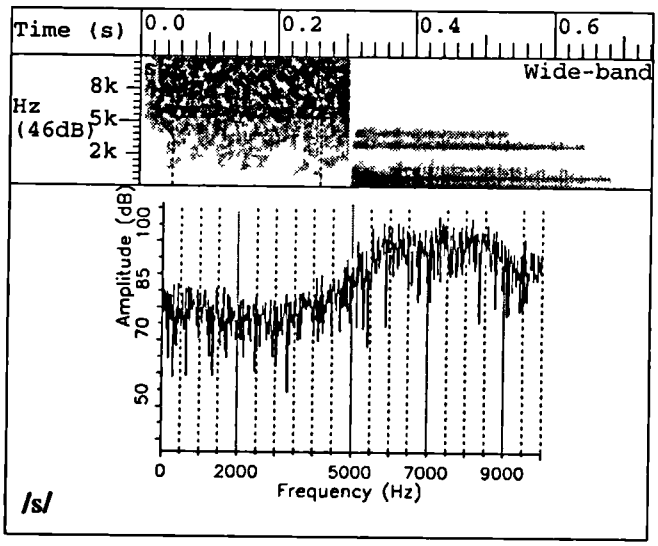

Time (s)
0.0
0.1
0.2
0.3
Hz
(43dB)
8k
5k
2k
Amplitude (dB)
100
85
70
50
0
2000
5000
7000
9000
Frequency (Hz)
/x/

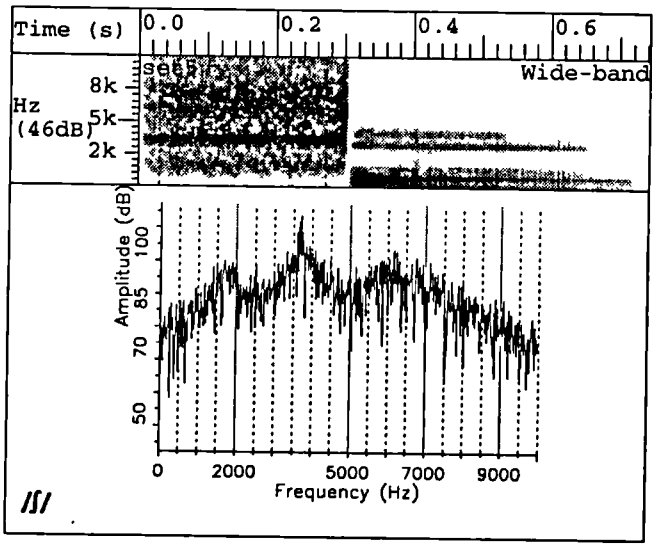

Time (s)
0.0
0.1
0.2
0.3
Hz
(41dB)
8k
5k
2k
Amplitude (dB)
100
85
70
50
0
2000
5000
7000
9000
Frequency (Hz)
/h/

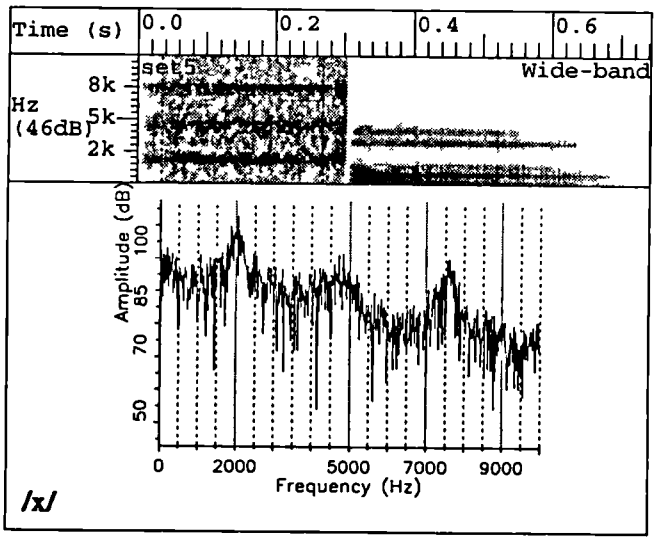
## LPC10a set

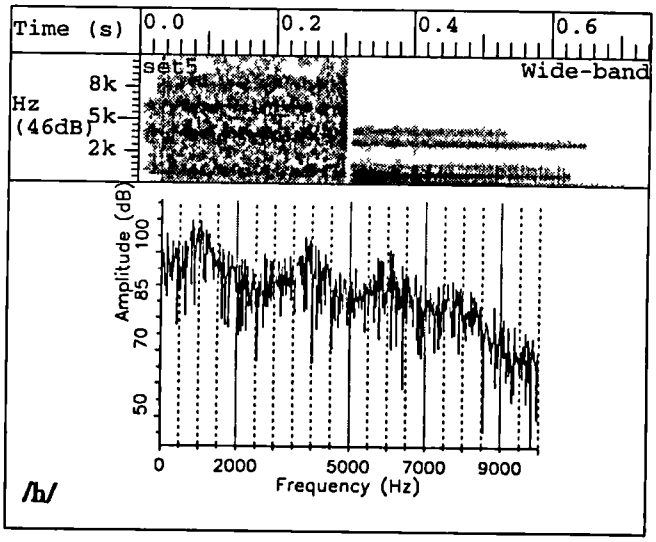

Time (s)
0.0
0.2
0.4
0.6
Wide-band
Hz
(46dB)
8k
5k
2k
Amplitude (dB)
100
85
70
50
0
2000
5000
7000
9000
Frequency (Hz)
/f/


Time (s)
0.0
0.2
0.4
0.6
Wide-band
Hz
(46dB)
8k
5k
2k
Amplitude (dB)
100
85
70
50
0
2000
5000
7000
9000
Frequency (Hz)
/θ/

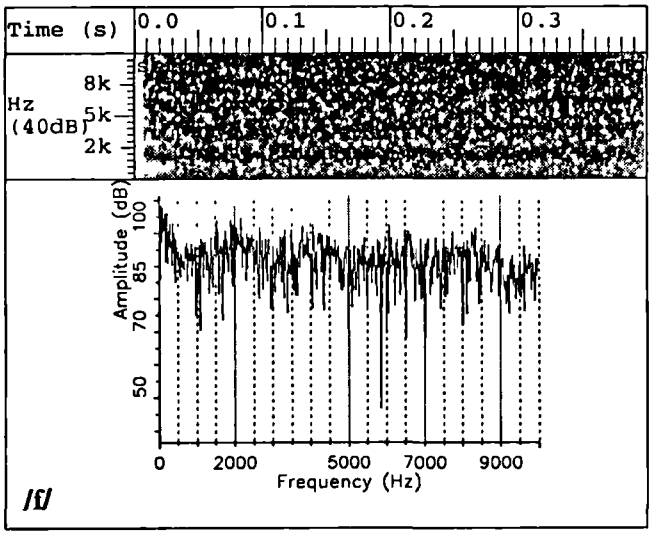

Time (s)
0.0
0.2
0.4
0.6
Wide-band
Hz
(46dB)
8k
5k
2k
Amplitude (dB)
100
85
70
50
0
2000
5000
7000
9000
Frequency (Hz)
/s/

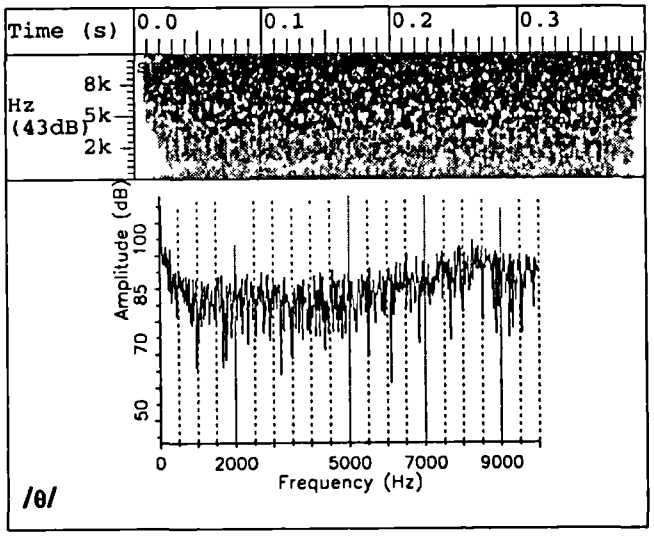

Time (s)
0.0
0.2
0.4
0.6
Wide-band
Hz
(46dB)
8k
5k
2k
Amplitude (dB)
100
85
70
50
0
2000
5000
7000
9000
Frequency (Hz)
/ʃ/

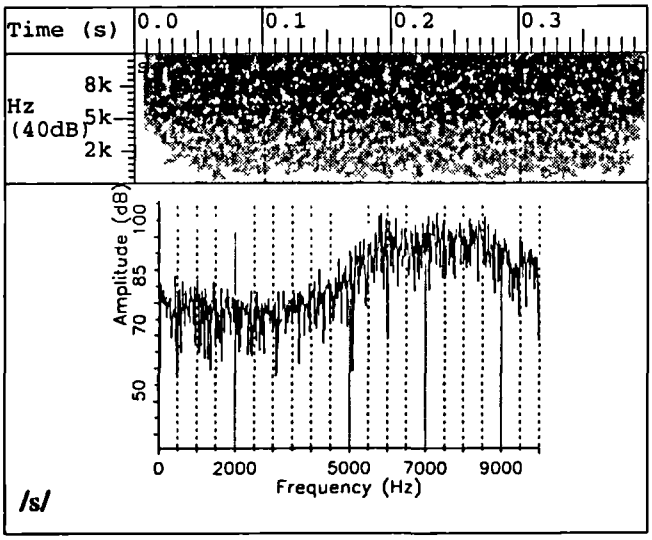

Time (s)
0.0
0.2
0.4
0.6
Wide-band
Hz
(46dB)
8k
5k
2k
Amplitude (dB)
100
85
70
50
0
2000
5000
7000
9000
Frequency (Hz)
/x/

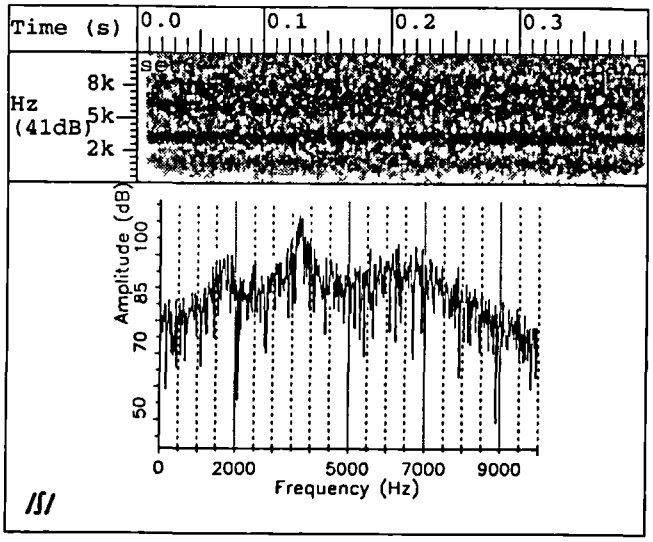

Time (s)
0.0
0.2
0.4
0.6
Wide-band
Hz
(46dB)
8k
5k
2k
Amplitude (dB)
100
85
70
50
0
2000
5000
7000
9000
Frequency (Hz)
/h/

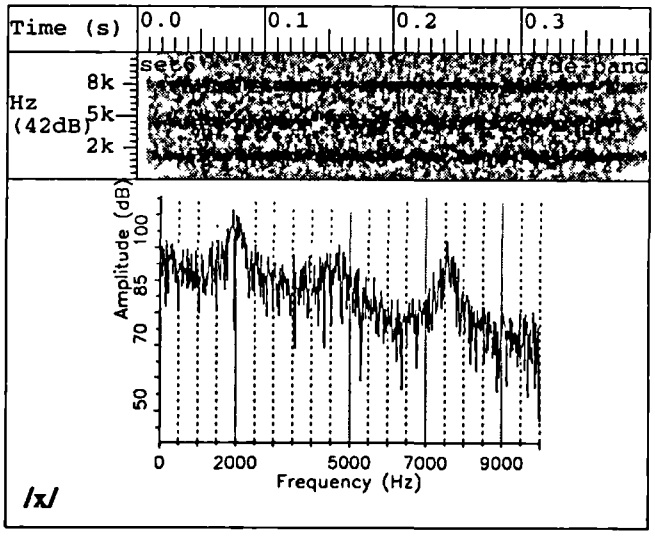
## LPC10 set

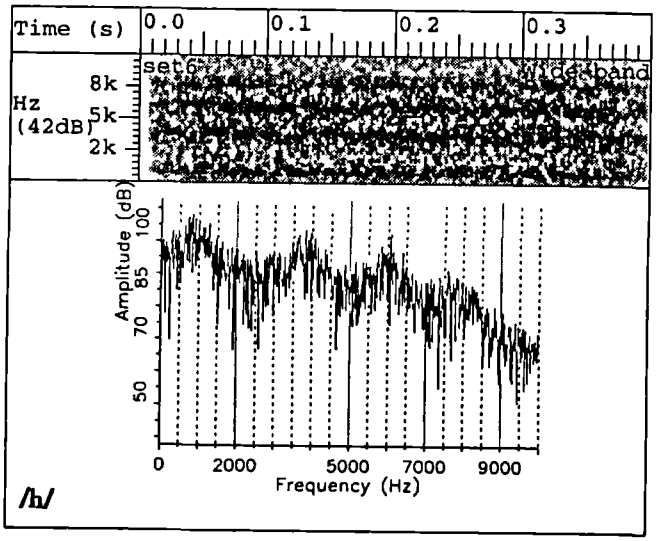

Time (s)
0.0
0.1
0.2
0.3
Hz
(40dB)
8k
5k
2k
Amplitude (dB)
100
85
70
50
0
2000
5000
7000
9000
Frequency (Hz)
/f/


Time (s)
0.0
0.1
0.2
0.3
Hz
(43dB)
8k
5k
2k
Amplitude (dB)
100
85
70
50
0
2000
5000
7000
9000
Frequency (Hz)
/θ/

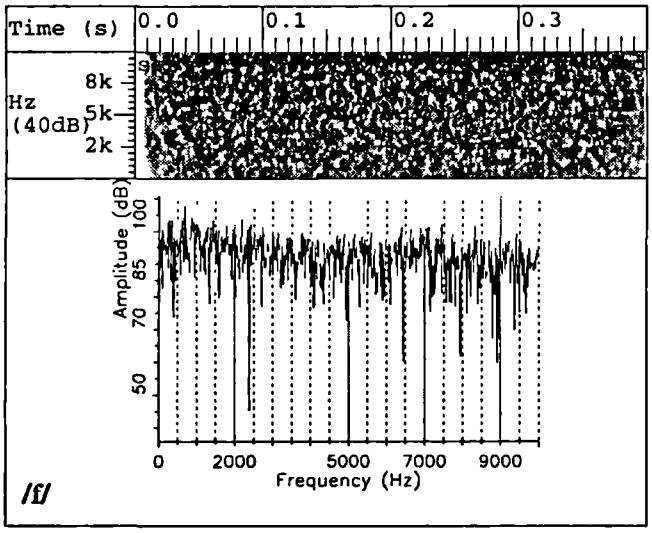

Time (s)
0.0
0.1
0.2
0.3
Hz
(40dB)
8k
5k
2k
Amplitude (dB)
100
85
70
50
0
2000
5000
7000
9000
Frequency (Hz)
/s/

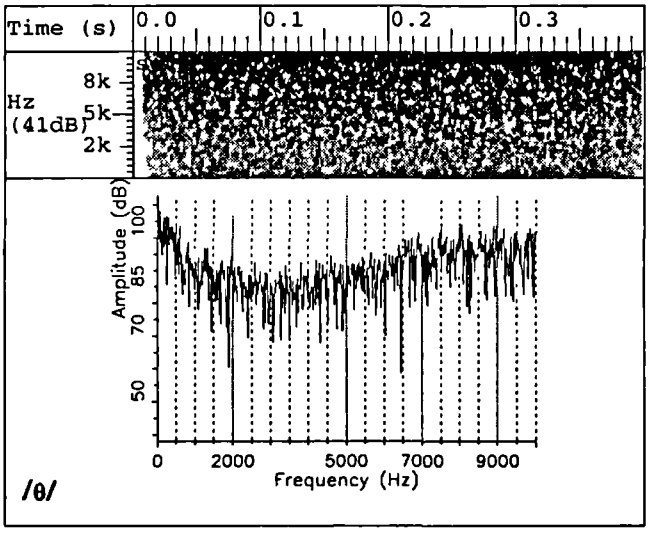

Time (s)
0.0
0.1
0.2
0.3
Hz
(41dB)
8k
5k
2k
Amplitude (dB)
100
85
70
50
0
2000
5000
7000
9000
Frequency (Hz)
/ʃ/

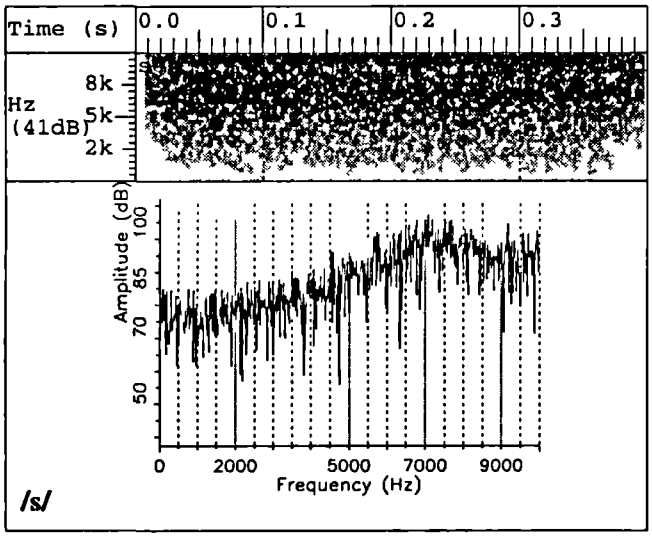

Time (s)
0.0
0.1
0.2
0.3
Hz
(42dB)
8k
5k
2k
Amplitude (dB)
100
85
70
50
0
2000
5000
7000
9000
Frequency (Hz)
/x/

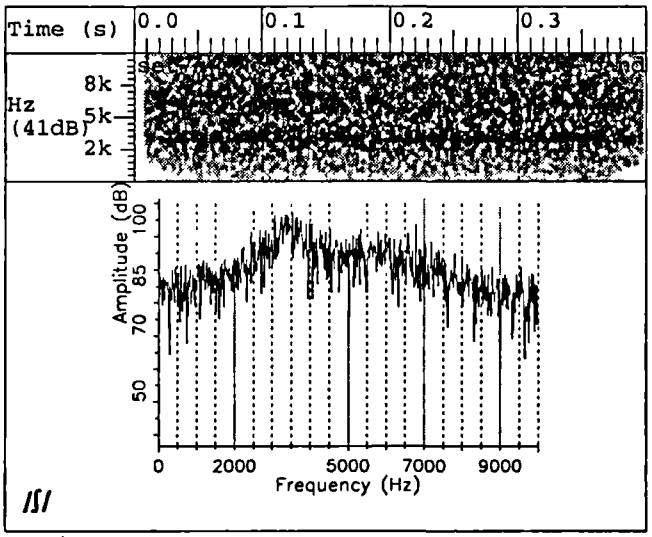

Time (s)
0.0
0.1
0.2
0.3
Hz
(42dB)
8k
5k
2k
Amplitude (dB)
100
85
70
50
0
2000
5000
7000
9000
Frequency (Hz)
/h/

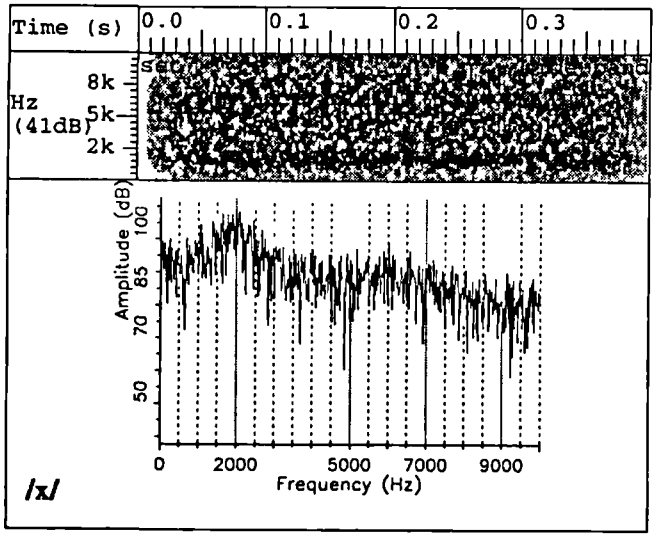
## LPC4 set

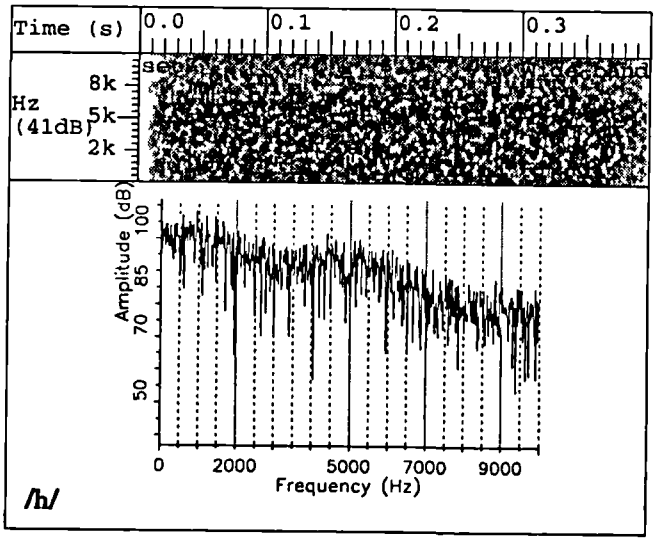

Time (s)
0.0
0.1
0.2
0.3
Hz
(40dB)
8k
5k
2k
Amplitude (dB)
100
85
70
50
0
2000
5000
7000
9000
Frequency (Hz)
/f/


Time (s)
0.0
0.1
0.2
0.3
Hz
(41dB)
8k
5k
2k
Amplitude (dB)
100
85
70
50
0
2000
5000
7000
9000
Frequency (Hz)
/θ/


Time (s)
0.0
0.1
0.2
0.3
Hz
(41dB)
8k
5k
2k
Amplitude (dB)
100
85
70
50
0
2000
5000
7000
9000
Frequency (Hz)
/s/


Time (s)
0.0
0.1
0.2
0.3
Hz
(41dB)
8k
5k
2k
Amplitude (dB)
100
85
70
50
0
2000
5000
7000
9000
Frequency (Hz)
/ʃ/


Time (s)
0.0
0.1
0.2
0.3
Hz
(41dB)
8k
5k
2k
Amplitude (dB)
100
85
70
50
0
2000
5000
7000
9000
Frequency (Hz)
/x/


Time (s)
0.0
0.1
0.2
0.3
Hz
(41dB)
8k
5k
2k
Amplitude (dB)
100
85
70
50
0
2000
5000
7000
9000
Frequency (Hz)
/h/